Network Science – Applications in Technology, Business and Social Media

Dissertation

zur Erlangung des akademischen Grades

doctor rerum politicarum

(Doktor der Wirtschaftswissenschaft)

eingereicht an der

Wirtschaftswissenschaftlichen Fakultät der

Humboldt-Universität zu Berlin

von

Annika Baumann, M. Sc.

Präsidentin der Humboldt-Universität zu Berlin:

Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Wirtschaftswissenschaftlichen Fakultät:

Prof. Dr. Christian Schade

Gutachter/in: 1. Prof. Dr. Stefan Lessmann

2. Prof. Dr. Hanna Krasnova

Datum des Kolloquiums: 04.04.2018

This thesis is dedicated to my parents and all people who supported me along the way.

Thank you.

Abstract

Networks constitute an integral part of our lives. Both in technical and social contexts, they can represent connections and relationships between entities of various kinds, enabling a deeper analysis of the underlying structure. One of the most important communication networks of recent times is the Internet. Increasing numbers of people use this worldwide network for manifold activities such as information searches, communication and commercial operations. Numerous technical applications, services and businesses are built based on this communication network and are therefore fully dependent on its functionality; for example, the World Wide Web, where social media and e-commerce websites belong, exerts a considerable influence on our everyday life and thus changes our way of thinking and acting. Additionally, the widespread propagation of handheld devices such as tablets and smartphones boosts the importance of the Internet by enabling individuals to access the mobile Internet, independently of location and time. In view of this, the emergence of the Internet has led to large changes in everyday life, which are examined in part in this dissertation.

Overall, the present dissertation is subdivided into three areas, which build hierarchically on each another and are based on the traditional three dimensions of information systems, comprising perspectives technology, management and organization.

At the core of this dissertation is the technological perspective, centered on an analysis of the Internet network using the mathematical-methodical aspect of graph theory. This work investigates the network and two selected applications that build on this communication network. The main goal is to derive insights regarding the general structure of the underlying network, in order to be able to understand its general communication functionality, detect potential bottlenecks and improve the susceptibility of the network to errors and deliberate attacks.

The second part of the thesis deals with the management perspective, which focuses on econometric considerations. In this case, another well-known application of the Internet network is the focus of attention: the World Wide Web, especially e-commerce websites which have a commercial implication. The focus lies on the understanding and prediction of user behavior, with the help of methods in the area of predictive modeling. A better anticipation of the behavior of customers in the online context enables companies to achieve higher revenues, through being able to make more informed business decisions.

The third area includes the organizational perspective, from the point of view of users of the Internet and its underlying applications. Here, two specific sub-areas are selected. The first area revolves around another type of World Wide Web application, social media websites, with the goal of understanding how sub-groups of users utilize these in different ways. The second area is centered around the aspect of how the propagation of mobile devices, and especially smartphones, in combination with pervasive Internet access, influences individuals in their personal and professional environments.

Based on these three perspectives, a total of 18 studies were conducted within the scope of this dissertation, using different methodological applications to gain scientific insights with respect to the areas examined.

Zusammenfassung

Netzwerke stellen einen integralen Bestandteil unseres Lebens dar. Sowohl im technischen als auch im sozialen Kontext können diese eingesetzt werden um Zusammenhänge und Beziehungen zwischen Entitäten verschiedener Klassen darzustellen und zu analysieren. Eines der wichtigsten Kommunikations-Netzwerke der jüngeren Zeit ist das Internet. Weltweit umspannend nutzen es immer mehr Menschen um verschiedensten Aktivitäten nachzugehen wie beispielweise der Informationssuche, der Kommunikation mit anderen und für Online-Geschäfte. Zahlreiche technische Anwendungen, Services und Unternehmen sind auf Grundlage dieses Kommunikationsnetzwerkes aufgebaut und entsprechend von dessen Funktionalität ganzheitlich abhängig. Beispielsweise übt das World Wide Web, zu dem auch die Sozialen Medien als auch E-Commerce Webseiten gehören, einen erheblichen Einfluss auf unser alltägliches Leben aus und verändert dadurch unser Denken und Handeln. Damit in Zusammenhang steht die starke Verbreitung von tragbaren Endgeräten wie Tablets und Smartphones welche die Möglichkeit offenbaren nahezu allgegenwärtigen Zugang zum Internet zu erhalten. Zusammengenommen führen diese Aspekte zu starken Veränderungen im Alltag welche in Teilaspekten in dieser Dissertation untersucht werden.

Insgesamt ist die vorliegende Dissertation in drei Bereiche unterteilt, welche hierarchisch aufeinander aufbauen und auf der traditionellen Perspektive der drei Dimensionen von Informationssystemen basieren welche die Technologie, das Management und die Organisation umfassen.

Im Zentrum der Dissertation steht hierbei die Technologie-Dimension in dessen Rahmen Netzwerke unter Nutzung des mathematisch-methodischen Aspekts der Graphentheorie analysiert werden. Hierbei werden das Internet-Netzwerk als auch Applikationen, die auf diesem Kommunikationsnetzwerk aufbauen, untersucht um deren Resilienz und Struktur besser zu verstehen um darauf basierend Ansätze zur Verbesserung der Fehleranfälligkeit und der Abwehr vorsätzlicher Angriffe abzuleiten.

Der zweite Teilbereich der vorliegenden Arbeit wechselt die Perspektive hin zum Management, in dem ökonomische Betrachtungen im Vordergrund stehen. Hierbei rückt das World Wide Web als eine der bekanntesten Anwendung des Internet-Netzwerks in das Zentrum. Konkret werden E-Commerce-Webseiten, die eine kommerzielle Implikation aufweisen, als Anwendungsbeispiel verwendet. Unter Nutzung von Methoden der prädikativen Modellierung stehen das bessere Verständnis und die Möglichkeit der Vorhersage von Nutzerverhalten im Fokus. Ein besseres Antizipieren des Verhaltens von Kunden im Internet unterstützt die Möglichkeit aus ökonomischer Sicht höhere Gewinne zu generieren, da die Management-Ebene dazu befähigt wird strategisch bessere Entscheidungen zu treffen.

Der dritte Bereich umfasst die Organisations-Perspektive aus Sicht der Nutzer, welche das Internet und dessen Applikationen anwenden. Hier wurden im Rahmen der Dissertation zwei spezielle Unterbereiche selektiert. Der erste Unterbereich betrachtet einen weiteren speziellen Bereich des World Wide Web welcher Webseiten Sozialer Medien umfasst und analysiert wie verschiedene Nutzergruppe diese verwenden. Der zweite Unterbereich befasst sich mit dem Einfluss der weitläufigen Verbreitung von mobilen Endgeräten in Kombination mit der damit verbundenen Möglichkeit des allgegenwärtigen mobilen Internetzugangs auf Aspekte des persönlichen und beruflichen Lebens von Individuen.

Aufbauend auf diesen drei Perspektiven wurden im Rahmen dieser Dissertation insgesamt 18 Studien durchgeführt, die sich unterschiedlicher methodischer Anwendungen bedienen um wissenschaftliche Erkenntnisse zu den vorgestellten Teilbereichen zu erlangen.

Table of Contents

Part 1:

Cumulative Dissertation

Publications Submitted with this Cumulative Dissertation

Publications in Journals

- Benjamin Fabian, Annika Baumann, Jessika Lackner (2015). “Topological Analysis of Cloud Service Connectivity”. Computers & Industrial Engineering 88, pp. 151-165. doi: 10.1016/j.cie.2015.06.009.
- Annika Baumann, Benjamin Fabian (2013). "Towards Measuring the Geographic and Political Resilience of the Internet". International Journal of Networking and Virtual Organisations 13(4), pp. 365-384. doi: 10.1504/IJNVO.2013.064465.
- Annika Baumann, Johannes Haupt, Fabian Gebert, Stefan Lessmann (2018). “Changing Perspectives: Using Graph Metrics to Predict Purchase Probabilities”. Expert Systems with Applications 94, pp. 137-148. doi: 10.1016/j.eswa.2017.10.046.
- Annika Baumann, Johannes Haupt, Fabian Gebert, Stefan Lessmann (2018). “The Price of Privacy: An Evaluation of the Economic Value of Collecting Clickstream Data”. Business & Information Systems Engineering. doi: 10.1007/s12599-018-0528-2.

Publications in Proceedings

- Robin Gubela, Stefan Lessmann, Johannes Haupt, Annika Baumann, Tillmann Radmer, Fabian Gebert (2017). “Revenue Uplift Modeling”. In: Proceedings of the 38th International Conference on Information Systems (ICIS’17), Seoul, South Korea, December 10th-13th.
- Olga Abramova, Annika Baumann, Hanna Krasnova, Stefan Lessmann (2017). “To Phub or not to Phub: Understanding Off-Task Smartphone Usage and its Consequences in the Academic Environment”. In: Proceedings of the 25th European Conference on Information Systems (ECIS’17), Guimarães, Portugal, June 5th-10th, pp. 1984-1999.
- Hanna Krasnova, Olga Abramova, Isabelle Notter, Annika Baumann (2016). “Why Phubbing is Toxic for Your Relationship: Understanding the Role of Smartphone Jealousy Among "Generation Y" Users”. In: Proceedings of the 24th European Conference on Information Systems (ECIS'16), Istanbul, Turkey, June 12th-15th.
- Olga Abramova, Annika Baumann, Hanna Krasnova, Peter Buxmann (2016). “Gender Differences in Online Dating: What Do We Know So Far? A Systematic Literature Review”. In: Hawaii International Conference on System Sciences (HICSS-49), Kauai, Hawaii, January 5th-8th. doi: 10.1109/HICSS.2016.481.
- Annika Baumann, Benjamin Fabian, Stefan Lessmann, Lars Holzberg (2016). “Twitter and the Political Landscape – A Graph Analysis of German Politicians”. Proceedings of the 24th European Conference on Information Systems (ECIS'16), Istanbul, Turkey, June 12th-15th.
- Annika Baumann, Benjamin Fabian (2015). “Vulnerability Against Internet Disruptions – A Graph-based Perspective”. In: Proceedings of the 10th International Conference on Critical Information Infrastructures Security (CRITIS’15), Berlin, Germany, October 5th-7th. doi: 10.1007/978-3-319-33331-1_10.
- Benjamin Fabian, Annika Baumann, Marian Keil (2015). “Privacy on Reddit? Towards Large-scale User Classification”. In: Proceedings of the 23rd European Conference on Information Systems (ECIS'15), Münster, Germany, May 26th-29th. doi: 10.18151/7217310.

- Annika Baumann, Stefan Lessmann, Kristof Coussement, Koen W. De Bock (2015). “Maximize What Matters: Predicting Customer Churn with Decision-centric Ensemble Selection”. In: Proceedings of the 23rd European Conference on Information Systems (ECIS'15), Münster, Germany, May 26th-29th. doi: 10.18151/7217273.
- Annika Baumann, Hanna Krasnova, Natasha Veltri, Yunsi Ye (2015). “Men, Women, Microblogging: Where Do We Stand?”. In: 12. Internationale Tagung Wirtschaftsinformatik (WI’15), Osnabrück, Germany, March 4th-6th.
- Annika Baumann, Benjamin Fabian (2014). “How Robust is the Internet? – Insights from Graph Analysis”. In: Proceedings of the 9th International Conference on Risks and Security of Internet and Systems (CRiSIS’14), Trento, Italy, August 27th-29th. doi: 10.1007/978-3-319-17127-2_18.
- Natasha Veltri, Hanna Krasnova, Annika Baumann, Neena Kalayamthanam (2014). “Gender Differences in Online Gaming: A Literature Review”. In: Twentieth Americas Conference on Information Systems (AMCIS’14), Savannah, USA, August 7th-9th.
- Annika Baumann, Benjamin Fabian (2014). “Who Runs the Internet? - Classifying Autonomous Systems into Industries”. In: 10th International Conference on Web Information Systems and Technologies – Volume 1: WEBIST, Barcelona, Spain, April 3rd-5th, ISBN 978-989-758-023-9, pp. 361-368. doi: 10.5220/0004936803610368
- Annika Baumann, Benjamin Fabian, Matthias Lischke (2014). “Exploring the Bitcoin Network”. In: Proceedings of the 10th International Conference on Web Information Systems and Technologies – Volume 1: WEBIST, April 3rd-5th, ISBN 978-989-758-023-9, pp. 369-374. doi: 10.5220/0004937303690374

Working Papers and Papers Being in Review Process

- Natasha Veltri, Hanna Krasnova, Annika Baumann. “Gender Differences in Blogging: Literature Review”. Working Paper.

In addition to the ones listed above, the following publications were also published in the course of my doctoral studies which are, however, not part of this dissertation:

- Tatiana Ermakova, Annika Baumann, Benjamin Fabian, Hanna Krasnova (2014). “Privacy Policies and Users’ Trust: Does Readability Matter?”. In: Twentieth Americas Conference on Information Systems (AMCIS’14), Savannah, USA, August 7th-9th.

Co-Authors

All papers submitted with this doctoral thesis were conducted in close collaboration with the co-authors mentioned below.

Professors

- Koen W. De Bock, IÉSEG School of Management, Lille (France)
- Peter Buxmann, Technische Universität Darmstadt, Darmstadt (Germany)
- Kristof Coussement, IÉSEG School of Management, Lille (France)
- Benjamin Fabian, Hochschule für Telekommunikation Leipzig, Leipzig (Germany)
- Hanna Krasnova, Universität Potsdam, Potsdam (Germany)
- Stefan Lessmann, Humboldt-Universität zu Berlin, Berlin (Germany)
- Natasha Veltri, University of Tamp, Florida (USA)

Research Associates

- Olga Abramova, Technische Universität Darmstadt, Darmstadt (Germany)
- Johannes Haupt, Humboldt-Universität zu Berlin, Berlin (Germany)

Students

- Robin Gubela, Humboldt-Universität zu Berlin, Berlin (Germany)
- Lars Holzberg, Humboldt-Universität zu Berlin, Berlin (Germany)
- Jessika Lackner, Humboldt-Universität zu Berlin, Berlin (Germany)
- Matthias Lischke, Humboldt-Universität zu Berlin, Berlin (Germany)
- Isabelle Notter, Universität Bern, Bern (Switzerland)
- Neena Kalayamthanam, Universität Bern, Bern (Switzerland)
- Marian Keil, Humboldt-Universität zu Berlin, Berlin (Germany)
- Tillmann Radmer, Humboldt-Universität zu Berlin, Berlin (Germany)
- Yunsi Ye, Humboldt-Universität zu Berlin, Berlin (Germany)

Others

- Fabian Gebert, Data Science Department, Akanoo GmbH, Hamburg (Germany)

1 Introduction, Motivation and Research Question

Networks are a common concept, and can be observed in everyday life. These can be, for example social, biological or technological networks, and can represent different entities forming some kind of relationship with each other. In the social context, for instance, they can show the friendship connections existing between individuals (Pitas, 2015) or in the case of biological networks can capture the interwoven net of neurons in the brain (Dorogovtsev and Mendes, 2013). One particular technological network has made a tremendous impact on business and society in the last decades, which is the Internet network. The emergence of the Internet into the public sphere in 1993 (Shepherd, 2007) entailed many changes to several aspects of the life of individuals. Today's Internet comprises a worldwide-spanning interconnected infrastructure, and is used by 3.8 billion people around the world (Statista, 2017a).

Prior to the existence of the Internet network, information needed either to be known or carefully sought from books, whilst today a vast stock of knowledge is only one click away. Furthermore, before the Internet emerged, letters were the written communication medium of choice; nowadays, e-mails can be used, reaching the destined recipient almost instantaneously. In 2015, around 205.6 billion e-mails were sent worldwide, and the number is expected to grow by five percent annually in the coming years (Radicati, 2015).

The Internet also changed the interaction of human beings in several more dimensions, such as the possibility of forming and maintaining friendships via social media websites. These kinds of websites exhibit tremendous usage statistics, with around 2.34 billion active users worldwide (Statista, 2017b). From a commercial perspective, changes are also apparent. Several services and industries depend fully on this technical infrastructure. E-commerce is a new business sector which emerged based upon the Internet infrastructure. The e-commerce sector is rapidly growing, reaching around $1.915 trillion of sales turnover in 2016 (eMarketer, 2016) with an increasing number of people browsing the Internet to shop online (Statista, 2017c). Some of the most important businesses with the highest market capitalization worldwide are online businesses, such as Alphabet Inc., Amazon.com and Facebook. All of these are completely dependent on the infrastructure the Internet provides (Gandel, 2016). These examples emphasize the modern importance of the Internet network for the personal and professional environment.

However, a complex infrastructure such as the Internet is susceptible to failures and deliberate attacks, and can suffer from certain bottlenecks, impairing its ability to stay connected und functional throughout. Several incidents such as natural disasters, power blackouts and accidental misconfigurations have demonstrated the Internet's vulnerability, causing a reduced communication ability possibly affecting wide geographic areas. Due to the importance of the Internet within business and society, this can have a tremendous negative impact. Although research into the Internet has been a focus for some years, certain aspects are still not fully understood due to the complex nature of the network. This thesis therefore strives to close several research gaps by analyzing the Internet infrastructure, its application and its users from various perspectives.

The main topics of this dissertation can be classified into the classical three dimensions of information systems, consisting of the technological, managerial and organizational perspectives (Laudon and Laudon, 2014, p.18). The technological perspective comprises the technical infrastructure of information systems including hardware, software, data storage and network components. Technlogy enables a organization and their management to accomplish their pre-defined business goals. The selection of appropriate technology components is part of the managements' responsibility and needs to be carried out carefully. By doing so, information technology enables the organization and its management to achieve financial stability through strategic business decisions which can be supported by information technology. The organizational dimension centers around the individuals who interact with information technology. Such individuals can be the organization as a whole or the individual people. The business is deeply interconnected with its information system structure causing a interwoven net of individuals, technology and the organization. By considering all these three dimensions, this thesis realizes a comprehensive understanding of all aspects of information systems in the selected areas.

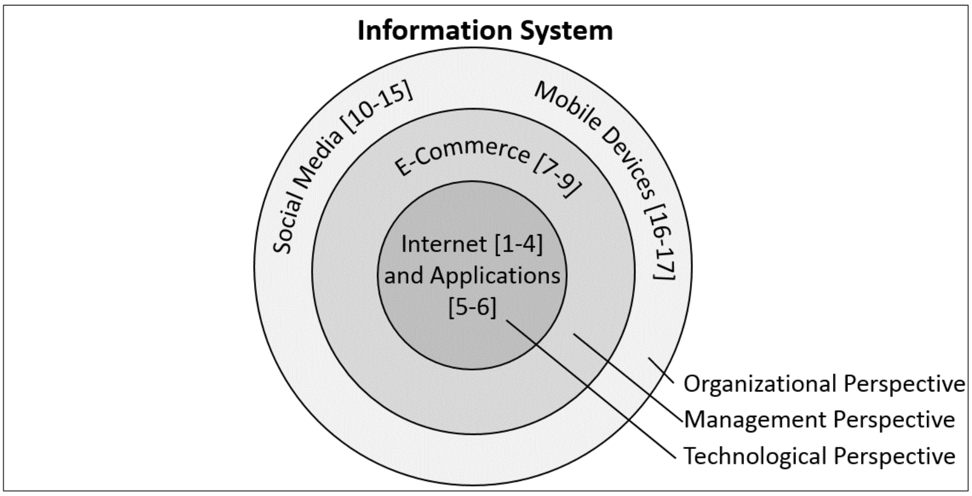

Figure 1. *Layer representation of the main topics considered in this dissertation. The numbers in brackets represent the relevant articles (see Section 3).*

In addition to the general structure of the Internet network, selected applications in technology, business and social media, which are based upon the Internet infrastructure, are the focus of this work. Figure 1 presents the main topics of this dissertation in a thematic context related to the three dimensions of information systems. The figure shows three different layers in the center. The core of the layer chart depicts the technological aspect of information systems, which constitutes the basis for all subsequent layers. The specific infrastructure in the case of this dissertation is the Internet network; its structure and robustness will be investigated. In addition to the Internet network in general, two different technological services are also the focus of this work: cloud computing and the cryptocurrency Bitcoin. Both services will be investigated from a technological viewpoint. In the case of Bitcoin, the underlying network of online payment transactions is explored, with the aim of understanding its general structure and composition. In the case of cloud computing, an understanding of the connectedness of different service providers to the Internet through analyzing their service availability is the goal.

On top of the Internet infrastructure, and therefore fully dependent on it, lies the management perspective. This layer focuses on the econometric perspective of the Internet. In this dissertation, the e-commerce sector is explored, and especially the understanding and prediction of the behavior of website visitors using clickstream data, with the aim of supporting managerial decision making and strengthening company success through improved decision making.

Following the technological and management perspectives, the analysis of aspects of the organizational perspective, focusing on the investigation of user behavior from various points of view, also forms part of this thesis. The outer layer of Figure 1 represents this dimension of information systems. In general, the two different use cases of social media and mobile devices are relevant, and in both cases, the understanding of users and their behavior is the focus. Social media platforms are based on the global Internet network, and mobile devices make it possible to access the Internet independently of location, causing an even more entangled and ubiquitously accessible communication network. These services therefore have a tremendous influence on the behavior of individuals.

In general, this dissertation aims to answer the following high-level research questions, each located at a different layer and from a different viewpoint:

RQ 1.1:	How robust is the Internet infrastructure, both in general and when considering country-based characteristics?
RQ 1.2:	What is the structure of the Bitcoin network?
RQ 1.3:	How well are cloud service providers connected to the Internet network?
RQ 2:	How can predictive modeling be used to estimate user behavior for e-commerce applications?
RQ 3.1:	What are the gender differences in social media applications, and how do specific communities use social media platforms?
RQ 3.2:	How does the ubiquity of smartphones affect individuals in personal and professional environments?

This dissertation uses a range of different methodological applications to answer the above research questions, as summarized in Figure 2. The main architectural concept of this dissertation are network-based structures. Graphs can represent these real-world networks. A graph is a mathematical notation whereby specific objects of interest, e.g. routers in the case of the Internet, or individuals in the case of social networks, form nodes which are connected via edges. Edges represent a certain interaction between two nodes, e.g. a data transfer in the case of the Internet, or an existing friendship between users of social media websites. These graphs can then be used to understand the specific phenomena and characteristics of each network. The methodological approach of graph analysis is utilized in research from the technological perspective and partly from the organizational perspective, in relation to social media analysis. The second methodological direction utilized in this thesis is predictive modeling, applied to the management perspective in the e-commerce context. This is a method which uses as its basis either structured or non-structured data, in order to find relationships and patterns in the information provided. The relations in the data are detected and learned by a statistical model. A huge variety of models exists, such as regression and classification algorithms, whose application depends on the outcome required from the data. These methodological applications are relevant in the e-commerce context, since the anticipation of user behavior can help in making more informed business decisions and generating higher revenue for companies.

Other methods applied in this dissertation are comprehensive literature surveys, used to systematically work up the current state of research in a particular area, in connection with relevant scientific theories, including the analysis of gender differences in social media applications such as blogs, microblogs and online dating platforms. Furthermore, in the case of the research area of mobile devices, a mixed method design is applied which consists of varying compositions of methodologies such as surveys, observations and a quasi-experimental setting, in order to understand how smartphones affect personal relationships and the behavior of individuals.

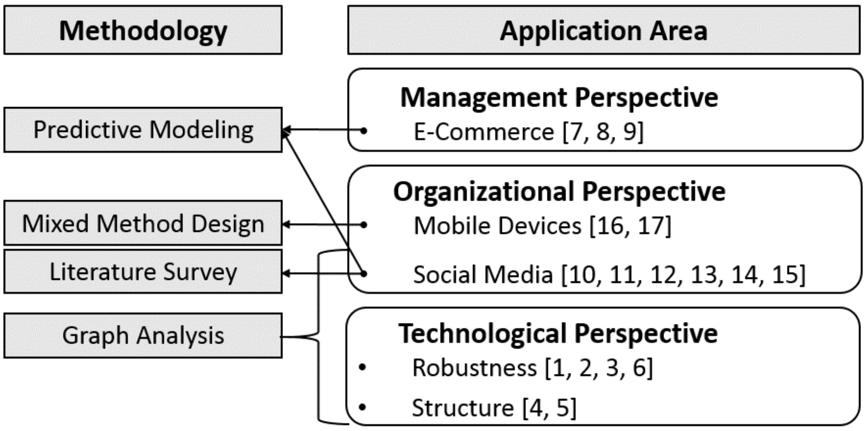

Figure 2. *Methodologies applied in the course of this dissertation. The numbers in brackets represent the relevant articles (see Section 3).*

The structure of the thesis is as follows: the second chapter will explain the specific research context for the thesis. This will be used to connect the topics included in this thesis at a higher level, and to present them in their overall research context. The third chapter will briefly summarize the papers submitted with this thesis. The fourth chapter concludes the first part of this dissertation with a descriptive view of published and submitted papers, in terms of their qualitative aspects with respect to three different rating systems.

2 Research Context

Central to this dissertation is an analysis of a network structure, i.e. the Internet, from several perspectives. Each perspective—technological, management or organizational—is itself built upon other fundamental concepts. This chapter introduces the main research context and concepts on which this dissertation is grounded, from a theoretical point of view. In doing so, the research topics of this dissertation are brought into their thematic context. While the first part focuses on graph theory and Internet research, the second and third parts concentrate on the World Wide Web applications of e-commerce and social media.

2.1 Technological Perspective: the Internet Network

The first dimension of information systems on which this dissertation is premised is the technological perspective. The different components of hardware, data storage techniques and communication technology, in combination with network structures, together comprise the IT infrastructure of organizations. While hardware consists of the physical components of electronic devices, centered around the concepts of input, information processing and output, data storage techniques are the physical components through which information can be stored digitally. Electronic devices can be connected through network structures using communication technology elements. At the core of this dissertation lies the Internet network, which consists of many thousands of connected electronic devices around the world, and the focus is on this specific element of the technological dimension of information systems.

The evolution of the Internet started in 1969 with the Advanced Research Projects Agency Network (ARPANET), which aimed to be a communication network which was highly resilient towards external factors (Tanenbaum, 2003, p.68). Within a few years, this initial network, built solely for research and military purposes and so far consisting of only a few entry points, grew rapidly. However, its commercial applications began in 1993 (Shepherd, 2007), and since then, the number of connection points has grown ever greater. Given the importance of the Internet today for social, communicational and commercial purposes, its resilience against deliberate attacks, accidents or failures has become critical for both businesses and society. This dissertation therefore aims to understand and give indications for improving the robustness of the Internet network. Next to the Internet itself, two specific Internet-based applications and their structure and robustness are at the center of this thesis; these are cloud computing applications and the cryptocurrency Bitcoin.

Over the course of its existence, the Internet as a worldwide-spanning network has fallen victim to several incidents of different categories, affecting its communication ability to various extents. In addition to general maintenance issues, several external impacts can also affect the structure of the Internet network negatively. Natural disasters such as hurricanes and earthquakes, accidents (e.g. cutting of underwater cables, power blackouts and misconfigurations) and deliberate attacks (e.g. Distributed Denial of Service Attacks and Internet worms) have harmed the Internet's communication ability in the past (Sterbenz et al., 2010; Wu et al., 2007). An understanding of its structure in order to locate potential bottlenecks and weak spots is therefore essential, since business and society depend largely on Internet-based services. In particular, applications whose revenue structure is entirely based on the Internet are affected in a negative way (ENISA, 2017).

Graphs can represent networks such as the Internet. They are a mathematical concept originating from the middle of the 18th century, and were the methodological approach used to solve the famous Königsberg problem (Barabási, 2016, p. 43). Graphs have been used for various applications and in all kinds of disciplines, for example biology, physics and information systems amongst others. The broad spread of possible application areas is an indicator of their high methodological relevance.

From a theoretical point of view, a graph $G = (V, E)$ consists of a set of nodes V which are pair-wise connected via edges E. The number of nodes composing a graph of is denoted as n and the number of edges as m. Edges in a network can be either directed or undirected, meaning that a single edge either points specifically in one direction (i.e. a directed graph) or does not (i.e. an undirected graph). Furthermore, each edge can have a label; the most commonly used label is a numerical weight, which assigns a specific value to a single edge in order to indicate, for example, the importance or load capacity of an edge. Graphs that include numerical edge labels are known as weighted graphs. A mathematical notation for representing a graph is the $n \times n$ adjacency matrix. In case of unweighted graphs, an individual value in the adjacency matrix is set to one if there is an edge between the two considered nodes, and zero if there is no edge between them. In case of weighted graphs, these values change to the respective weight of the edge that connects the two considered nodes.

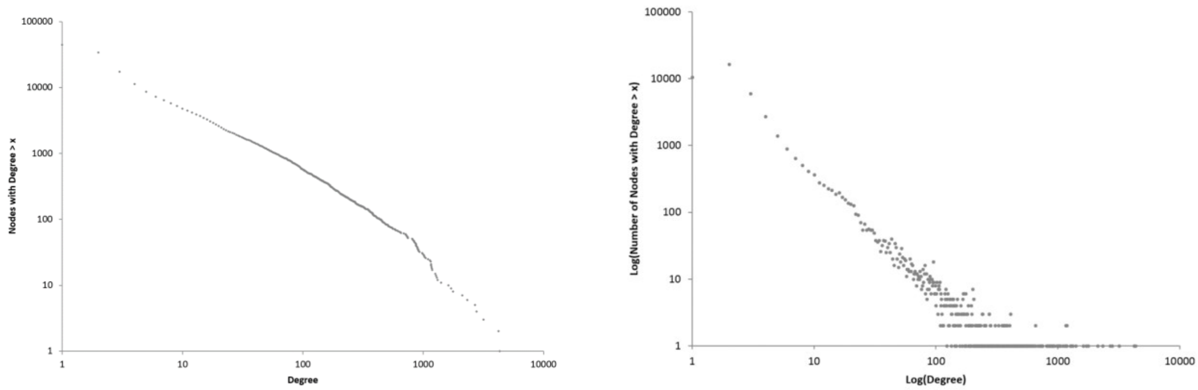

Figure 3. The resultant linear degree distribution (left) and on a log-log scale (right) of the Internet dataset used in this research (Baumann, 2013).

Various metrics exist for characterizing the structure and composition of graphs. Some commonly used metrics in research are based either on distance measures in the network or on the centrality of graph elements. One of the most popular of these is the degree centrality (or short degree), which constitutes the number of edges of a node. In the case of undirected graphs, the average degree in the network can be calculated using $< k >= \frac{1}{n} \sum_{i=1}^{n} k_i$, where n is the number of nodes in the network and k_i the degree of node i. In the case of directed graphs, there is a distinction between the in- and the out-degree, the former measuring the number of incoming edges, and the latter the number of outgoing links from a node. The distribution of degrees in the Internet network resembles a power-law distribution, meaning that there are many nodes with a low degree but few with an extremely high degree (see Figure 3). Other

commonly used metrics based on the centrality concept are closeness centrality, eigenvector centrality and betweenness centrality, all of which measure the centrality of single nodes or the whole network, from varying points of view.
A commonly used distance-based metric is the shortest path length, which measures the average number of hops necessary for a node in the network to reach every other node. The metric indicating the average shortest path length for all nodes in the network can be calculated via $l_{Graph} = \frac{1}{n(n-1)} \sum_{i \neq j} d(v_i, v_j)$, where n is the number of nodes in the network and $d(v_i, v_j)$ the minimum number of hops between nodes v_i and v_j. The node-based view of the average shortest path length is the eccentricity, which measures the longest shortest path in the graph for any node in the network. Derived from the eccentricity are the metrics of diameter and radius; the former measures the longest shortest path for each possible distinct pair of source and destination nodes in the graph, and the latter indicates its lowest value in the graph.
Several further metrics are available in the literature, each with a specific view of the graph structure within a specific context. Mahadevan et al. (2006) provide a comprehensive overview of the most commonly used metrics in Internet graph research.
Graphs can either be based on real-life data or artificially created models. Two of the most well-known models that have been used in research as a means to represent the Internet infrastructure are the Erdös-Renyi model (ER model; Erdös and Rényi, 1959) and the Barabási-Albert model (BA model; Barabási and Albert, 1999). Each of these exhibits specific characteristics. The ER model starts with a number of nodes n and a fixed connection probability p between two nodes, in the range $[0,1]$. In contrast, the BA model algorithm dynamically changes the likelihood p of connection in such a way that a preferential attachment is considered; that is, nodes with a high degree are preferred as connection partners for a new node in the network.
Both models exhibit a small-world character, as does the real Internet topology. This is a concept coined by Travers and Milgram (1967), based on an experiment where people in the US cities of Omaha and Nebraska were given letters, which they had to hand over personally to other people whom they thought would be suitable to get the letter to its destination in Boston, Massachusetts. On average, six delivery steps were necessary to solve this task. Applied in the context of graphs, this means that each node can be reached by any other node in the network using only a few steps. In addition to this small-world character, the BA model has the characteristic of a power-law distribution with respect to the structure of edges in the network. This means that there are many nodes in the network which have only a small number of connections, whereas there are few nodes with a high number of connections. Upon deeper investigation of the structure of the real Internet network, it became clear that in addition to a small-world character, it also exhibits a power-law distribution (Faloutsos et al., 1999). Networks which exhibit a degree distribution matching a power law are known as scale-free (Barabási, 2016).
The dynamic character of the Internet makes the task of mapping the network as a graph rather challenging, since its topology and routing paths change on a daily basis. An analysis of a current snapshot or a dynamic map of the Internet structure is therefore crucial. Here, the modeling of the Internet and its connection points can be done by focusing on different granularities, as shown in Figure 4. The most fine-grained representation form is at the level of IP addresses, where each IP address represents a node in the graph. Several IP addresses can

belong to one specific router and can therefore be aggregated into a single entity. This constitutes the second representation option on the level of routers which is less detailed compared to the level of IP-addresses. Based on geographic proximity, the point-of-presence level uses a single node for several routers within a specific close proximity, such as being located in the same building or area. Another common option for a less fine-grained form of representation is the modeling of the Internet graph based upon autonomous systems. An autonomous system consists of a number of routers that belong to the same technical administration entity (Hawkinson and Bates, 1996). This last option is chosen for the research in this thesis with respect to the investigation of the robustness and structure of the Internet. This is due to its advantage in that since each autonomous system consists of a set of routers, the general structure of the Internet network is still captured, while at the same time the volatility of structure is minimized due to the high-level view. Furthermore, the calculation of computationally expensive metrics remains manageable within a reasonable period which might be especially crucial in case of real-time applications.

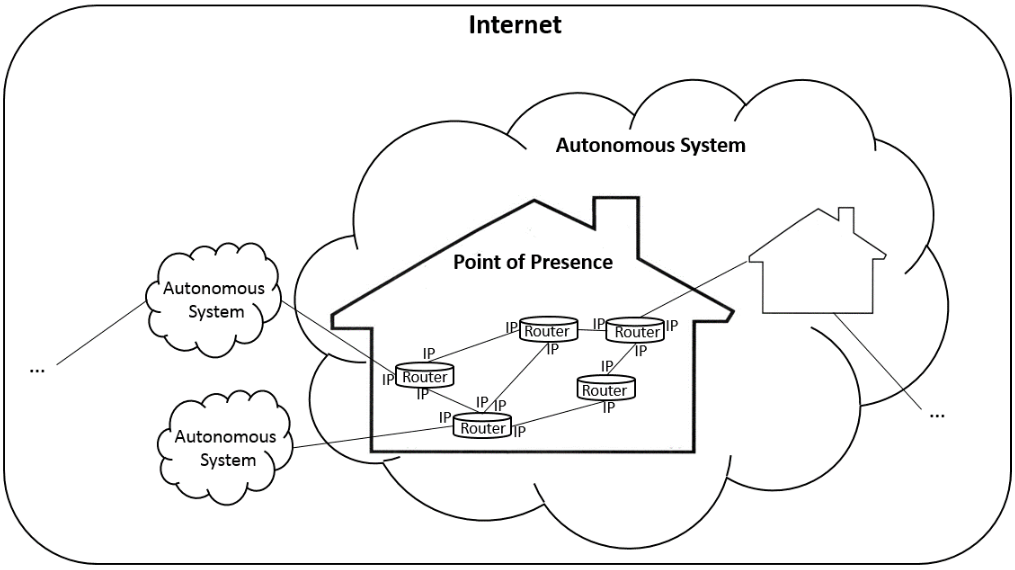

Figure 4. Different granularities representing the Internet as a graph structure.

A frequently used approach to measure the resilience of a graph structure is to successively remove graph elements from the network based on pre-defined characteristics. Here, failures in the graph structure are modeled using the random deletion of a graph element from the network, whilst malicious activities such as attacks target selected nodes in the network that are important in terms of communication ability. The order of the nodes removed from the network is mainly based on a list of nodes ranked according to a specific graph metric that captures the current characteristic of interest.

Robustness research on the Internet network started with artificially created models based upon the ER and BA models introduced above. Particularly in the early stages of Internet research, and due to the non-availability of reliable data sources, these models have been used to represent specific characteristics of the real-world network in order to capture its behavior and characteristics and to understand how the network reacts to random failures and targeted attacks. With a focus on Internet resilience, Albert et al. (2000) investigate the behavior of the

ER and BA models following the random and targeted removal of nodes from the network. Based on diameter as a quality measure of connectedness, the random removal of nodes affects the network to almost no extent, whereas targeted deletions based upon an ordered list of highest degree nodes disconnect the network into several components almost immediately. Cohen et al. (2000, 2001) and Crucitti et al. (2003, 2004) show similar results in their research.
In the following stages of research the focus was on Internet topology built upon real data taken from publically available mapping projects such as Caida's AS Rank (Caida AS Rank, 2012), Oregon RouteViews (RouteViews, 2013) and the dataset provided by the University of California in Los Angeles (UCLA, 2013), in order to generate a more accurate Internet network structure. Magoni (2003) investigates the robustness of the Internet topology at the router level. Even at a more fine-grained scale, the results are the same: the topology of the Internet is robust with respect to random failures, but is highly vulnerable to targeted attacks. Other approaches to obtaining a deeper understanding of the Internet's robustness have taken the direction of more complex graph element removal strategies, such as a combination of random and targeted elimination of graph elements and the consideration of other metrics (Park et al., 2003).
More recent work has collected real-world data in view of the economically driven routing character of the Internet network, i.e. policy-driven routing. As stated by Gao (2001, p.733), "connectivity does not imply reachability," meaning that policy-driven routing limits the possibility of available routing paths due to the underlying business relationships between autonomous systems. As shown by Dolev et al. (2006) and Kurant et al. (2007), the consideration of restricted access to routing possibilities makes the Internet network even more vulnerable in terms of targeted eliminations of graph elements, whereas the random removal of around 95 percent of the nodes in the network has only negligible effects on connectedness. Wu et al. (2007) state that policy-driven routing creates an effect where available alternative routing options cannot be used in some cases, thereby making the Internet network more vulnerable. Furthermore, Xiao et al. (2008) consider a more realistic attack strategy where only information about the graph elements in the local neighborhood of a node is available to a potential attacker. Starting from a number n of initial target elements in the network, the next targets are chosen from the direct neighborhood based upon certain selection criteria. Although not as destructive as the simple attack mode considered in prior literature, this strategy is shown to be reasonably effective, destroying a network's communication ability rather quickly.
The next research phase was characterized by the proposal of strategies and frameworks aiming to make the structure of the Internet more resilient towards external disturbances. While Schneider et al. (2011) hypothetically interchange edges in the network to create resilience, Sterbenz et al. (2010, 2011) and Smith et al. (2011) propose a framework which formally explains how to reduce, detect and deal with instabilities in the network structure.
In general, two main research streams can be identified. The first is the investigation of the robustness of the Internet from a graph perspective via graph element removal, using different datasets and strategies and moving towards more realistic approaches to modeling the topology and external disruptions. The second research stream is based upon the first, and proposes countermeasures to increase the resilience of the Internet network. Overall, research shows that due to the power-law characteristic of the topology of the Internet, failures affecting a random node in the network have almost no harmful effect on the network's communication ability.

Reasons are that the likelihood of affecting a central node is rather low, leading to potential routing alternatives available throughout the network. However, deliberate attacks which specifically target important nodes in the network that are highly relevant for communication ability may cause a sharply reduced capability for routing traffic through the network.

This dissertation adds insights into the robustness of the Internet's topology, based upon a recent set of publically available datasets which integrates and combines three different recent data sources. The final dataset consists of data taken from Caida AS Rank (Caida AS Rank, 2012), the UCLA dataset (UCLA, 2013), the Archipelago project of Caida (Caida Ark, 2013) and Internet Routing Registry Data (IRR, 2013). This allows the derivation of a reasonably comprehensive map of the topology of the Internet. However, due to the incompleteness of each data source, which arises from the complex nature of the Internet's topology, a "ground truth" of its network structure is not available, and this is an ongoing research topic (Oliveira et al., 2010, p.1). The Internet graph is then analyzed with the help of network theory, to derive the characteristics of its general structure based upon a rich set of different graph metrics. Various failure and attack strategies are then applied to the Internet graph to understand its resilience against these scenarios. Here, strategies commonly used in the literature are applied in order to understand the resilience of the derived current Internet topology (article [1]). Based upon these insights, a ranking system using the connectivity risk score is proposed, based upon a set of graph metrics relevant to the propensity of single nodes to have connectedness to the network (article [2]). This robustness research is then further extended by adopting a geographic and political perspective, using a country-based classification of autonomous systems with respect to resilience within geographic boundaries (article [3]). Furthermore, the autonomous systems forming the Internet network are classified into their respective industries in order to gather insights into the various players forming the network structure (article [4]).

Numerous applications are based upon the technology of the Internet. A subset of these are the focus of this dissertation, including the World Wide Web (i.e. e-commerce applications and social media websites; see Section 2.2), cloud computing and cryptocurrencies.

Cloud computing is an online service which offers IT infrastructure that is easily accessible through the Internet, independent of location and time aspects. Due to its nature, the service is only available when a working Internet connection is available, both on the client and the server sides. An extensive connection to the Internet network is therefore crucial in order to be independent of single node and edge outages. This dissertation examines the server side of cloud computing service connectivity, and analyzes how well autonomous systems are connected to the Internet from a graph-based perspective (article [5]).

Cryptocurrencies are a rather new concept of online currencies, and have an underlying cryptographic structure. The cryptocurrency occupying the highest market share and forming the focus of this research is Bitcoin. Bitcoin uses a Blockchain mechanism to verify transactions. This Blockchain information can be used to generate the transaction graph, which can then be analyzed in terms of several aspects. So far, main directions of research have been the analysis of the structure, functionality and the anonymity of users of the Bitcoin network (Morisse, 2015). The research in this dissertation further investigates the network structure of the transaction graph of the cryptocurrency Bitcoin up to October 2013 (article [6]).

2.2 Management Perspective: E-Commerce

The second main area of this dissertation deals with the management perspective of information systems. A company must operate under a certain degree of uncertainty, whereby decisions have to be made and business strategies need to be set. Here, the management level is responsible for resource allocation and decision making, in order to lead the company to financial success. The business area at the center of this thesis is the e-commerce sector, which is built upon the infrastructure of the Internet. E-commerce is an abbreviation for electronic commerce, which encompasses websites belonging to the commercial area of the Internet where products and services are sold (Cebi, 2013). Depending on the type of seller or buyer, those websites can be either Business-to-Business (B2B), Business-to-Customer (B2C) or Customer-to-Customer (C2C). In the case of B2B markets, both parties involved are businesses, while in case of B2C the business sells products or services to private individuals. In addition, individuals can use specific e-commerce platforms to sell their commodities, and these belong to the category C2C. In this thesis, commercial online platforms that are part of the B2C sector are considered.

The e-commerce sector is a competitive environment, with increasing numbers of participants entering the market each year. Accurate numbers are hard to estimate, but calculations based upon one set of assumptions state that there seem to be between 12 and 24 million online stores worldwide (Rachamin, 2014). However, the sales distribution among those stores seems to follow a power-law, meaning that only around 650,000 stores reach an annual sales amount of at least $1,000 (Rachamin, 2014). Therefore, making the right decisions at the right moment is crucial for business success, in order to outperform competitors and reach a sufficient annual sales volume. This is especially important since the e-commerce environment is highly dynamic and fast moving. Here, management actions have the aim of winning new customers and strengthening the relationship with existing customers in order to increase revenue for the e-business. However, due to uncertainty, it is often unclear which decisions are the best to make.

Compared to traditional brick-and-mortar shops, the e-commerce area differs in certain respects. Business transactions in traditional shops lead to direct customer interactions; this can achieve customer trust and satisfaction, which may have a positive impact on the purchase intention of an individual (Gefen and Straub, 2004). However, direct interactions with customers in the e-commerce context are limited; the only possible communication is via the website itself and through any customer service offered via e-mail or telephone. Furthermore, in the case of brick-and-mortar shops, the haptics of products can be instantaneously experienced, which is also an important determinant leading to a purchase decision (Peck and Childers, 2003). Additionally, this direct presence of the buyer and seller in one place has the effect that the purchasing process can be carried out immediately, through the simultaneous exchange of money and products, thereby representing a safe environment. In the e-commerce setting, this can only be achieved to a limited extent, and trust has to be given in advance. Interaction with potential buyers take place through the e-commerce website, where the real products sold have to be presented in a digital way with the help of media such as text, images or video, and the quality aspects of the product are therefore not directly experienced. Furthermore, the completion of the buying process is changed by the need for shipment of the bought product(s), which can lead to an initial reduction in trust towards the online shop.

In general, an increase in revenue can be achieved through various mechanisms such as new customers buying from the seller, re-purchases from existing customers and reducing the churn of existing customers through retention programs, possibly leading to more follow-up purchases. Two customer actions are therefore responsible for creating higher revenue: purchase behavior and re-purchase behavior; the latter is centered around the prevention of churn of existing customers. Several determinants have been identified through research that influence the customer's intention to (re-)purchase in the e-commerce context. In addition to objective factors such as the price and quality of a product (Liao and Cheung, 2001) and attributes of the e-commerce website and shop owner (Kuan et al., 2008; Jarvenpaa et al., 2000), the perceptions of customers with regard to trust and perceived risk towards the e-commerce shop and customer satisfaction (Kim et al., 2009) are essential.

Offering the visitor a website with a professional interface and maintaining a positive reputation is therefore helpful in increasing revenue. Furthermore, social presence is a determinant which forms trust (Gefen and Straub, 2004). However, since direct customer interactions are missing, other mechanisms must be established to strengthen the trust and satisfaction of website visitors in the e-commerce context. These mechanisms could, for example, be established through personalized digital interaction with the user (Gefen and Straub, 2004), such as providing customer service or recommendations for products for the user. However, giving unsuitable recommendations or disturbing a visitor with unwanted customer service mechanisms may be harmful, and may prevent a website visitor from becoming a buyer. Making the right decisions at the right time is therefore crucial. However, in the e-commerce context, a website may have an unmanageable amount of anonymous visitors at the same time, and making personalized decisions manually for each individual website visitor is not feasible.

In this regard, methods of predictive modeling have been applied in the e-commerce context which aim to understand user behavior better or to automatically anticipate how the user will behave in the future. In this way, they support management in making more informed business decisions in the e-commerce context by giving indications about user behaviors which can be used to align management decisions and actions accordingly. In general, methods of predictive modeling consist of three different elements: these are features built upon available data, an algorithm which learns relationships based upon the information input, and a response which represents the variable of interest.

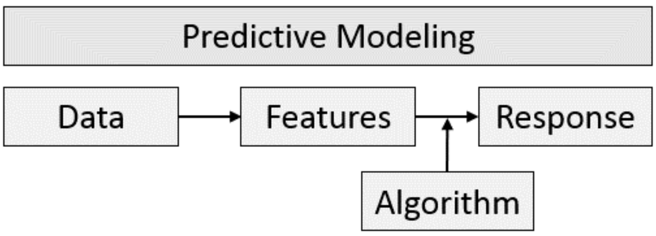

Figure 5. A generalized predictive modeling process.

To be able to predict the response variable of interest, an algorithmic model needs information it can process to learn specific patterns in the data leading to the specific outcome. In the case of the e-commerce context, this information can be gathered from various sources. For example, customer data can be used which is collected over the period of the membership or contractual duration of the customer. This can consist of information such as demographic data, membership or contract information and transaction data. However, the situation is more complicated if the website visitor is not known. In the traditional shopping environment, specific gestures of a customer can hint at his or her intention towards visiting a shop, while these markers are missing in the e-commerce context. Uncertainty is even higher as these anonymous customers can be identified only by an IP address. However, website visitors leave trace marks known as web logs or clickstreams, which represent the second major data source for the application of predictive modeling techniques in the e-commerce context. A clickstream can be defined as "a record of a user's activity on the Internet, including every web site and every page of every web site that the user visits, how long the user was on a page or site, in what order the pages were visited, among others things" (Garrie and Wong, 2007, p.131). The order of webpages visited, the time spent and the interaction with the website stored in the clickstream data can therefore be used to understand and anticipate the behavior of website visitors. A famous feature concept based upon clickstream data is recency, frequency and monetary (RFM) value analysis (Zhang et al., 2015), which reflects different dimensions of a customer's journey when visiting a website. Additionally, both data sources (traditional customer data and clickstream data) can be combined to obtain an even more complete picture of individual customers for the prediction task. The available information sources form the basis for deriving features (or covariates), i.e. specific bits of information in the data stream, which the algorithm can use for learning patterns and relationships in the data, leading to specific outcomes.

The selection of an appropriate algorithm constitutes the next step of predictive modeling. In general, such algorithms can be divided into supervised and unsupervised learning. In the case of supervised learning, data is available in which the outcome of interest is already known. The applied algorithmic model can then use this known information to detect and learn patterns in the data, in order to be able to predict outcomes for unknown data. In case of unsupervised learning, no data is available in which the true outcome is known, and the algorithm therefore needs to learn underlying patterns independently of the respective response variable. Unsupervised methods are applied, for example, in the area of customer segmentation, where individuals are separated into distinct, unknown groups that minimize inter-group similarity while maximizing intra-group similarity (Woo et al., 2005). The methods used in this dissertation focus on approaches involving supervised learning. The type of algorithmic method to use depends on the response variable, i.e. the outcome for prediction, which can be either a class label consisting of likelihoods of class affiliation or a continuous value; the former refers to classification, and the latter to regression methods. In the case of classification, a group membership is therefore of interest, whilst in the regression case, the estimation of a specific value is the focus. In this dissertation, classification algorithms are utilized. Classification algorithms commonly applied in research include decision trees, neural networks, support vector machines, instance-based algorithms and statistical learning algorithms (Kotsiantis, 2007).

Due to the variety of algorithmic models from which to select, and the combination of parameters that are adjustable for each model, benchmarking studies are used to comprehensively compare a selection of different models for a fixed setting. Depending on the response variable of interest and the data available, different models may be more applicable to specific settings. A systematic analysis of a selection of models for a single response type and a variety of datasets is therefore helpful in understanding specific prediction tasks.

Predictive models have been commonly used in e-commerce research to predict a variety of response variables. One type of response variable using classification methods is the division of website visitors into groups according to their intention of website visit (e.g. Moe, 2003). Marketing-related activities such as estimating the likelihood of banner clicks (e.g. Nottorf and Funk, 2013), customer churn (e.g. Moertini and Ibrahim, 2015) or personalization approaches such as displaying incentives (e.g. Pai et al., 2014) have also been considered. However, the most common response variable in the literature is the prediction of purchase behavior and conversion (e.g. Buckinx and Van den Poel, 2005). In this dissertation, two different response variables are in the focus: churn and purchase behavior.

As discussed above, one factor leading to reduced revenue for e-commerce shops is the churn in customers. An understanding of when a customer is likely to churn is therefore crucial for businesses, since retaining existing consumers is easier and more cost-effective than winning new ones through marketing activities (Bhattacharya, 1998; Colgate and Danaher, 2000). Even if loyalty programs exist (Kopalle et al., 2012), the churn behavior of customers constitutes still a major threat to several companies (Schweidel et al., 2008). This is especially true in the fast-moving digital age, where the cancellation of a membership can often be accomplished through only a few clicks. Retention programs are therefore crucial, but for the purposes of business revenue these should be personalized by only addressing actions to those customers who actually intend to stop buying from the company. Predictive models can help in understanding which customers are likely to churn. The prediction of churn is a binary classification problem, where customers fall into categories of either churning or non-churning. The applied algorithm assigns to each customer a likelihood reflecting the probability of individual churn behavior. This likelihood is then used to rank customers from highest to lowest churn probability, and those reaching a specific threshold are assigned to the churn behavior category. In this dissertation, a benchmark for the prediction of churn behavior in the telecommunication sector is applied in combination with a new proposed approach which outperforms all other benchmark models applied (see Section 3.2, article [7]).

The second response variable focused on in this dissertation is the prediction of purchase behavior, which constitutes the main driver for the creation of revenue for e-businesses. Understanding which customers are likely to buy can be helpful in understanding which factors are important in leading to purchase, and in potentially targeting with incentives those customers who are not yet likely to become buyers (Pai et al., 2014). This dissertation focuses on two specific aspects in the area of purchase prediction in the e-commerce context; these are privacy-related issues and the potential of graph-based features as a means of predicting user purchase behavior.

Privacy "is the ability to manage information about oneself" (Belanger et al., 2002, p. 249) and constitutes an important determinant of the perception of trustworthiness towards an e-commerce shop (Metzger, 2004). Trust, amongst other aspects, is itself an important antecedent of purchase intention, which may lead to a simple website visitor turning into a buyer (Gefen and Straub, 2004). As stated above, in e-commerce, personal information is collected in several ways. One of these requires the direct action and consent of the website visitor, who willingly provides personal information in order to complete a purchase process. The handling of such data is regulated in the terms of service of the e-commerce shop. Another method of data collection is the gathering of clickstream data; this takes place without the direct consent of the website visitor, who may prefer to remain anonymous. However, clickstream data can then be used to re-identify such users via behavioral indicators (O'Connell and Walker, 2014; Yang, 2010) and can therefore constitute highly sensitive information. However, clickstream data information can also be a basis for providing visitors with a more enjoyable experience on the website, through personalized content, for example. There is therefore a trade-off between the collection of potentially privacy-sensitive information and the desire of users to stay anonymous. The collection of user data constitutes a threat to the user's privacy, which needs to be considered when carrying out predictive modeling in the e-commerce context. Thus far, the privacy aspect of predictive modeling has been firmly in the center of research. Padmanabhan et al. (2006) focus on the collection of clickstreams over several websites, and Stange and Funk (2015) analyze the sample size needed to achieve reasonably accurate prediction results. However, the analysis of single-site clickstream data with respect to which features are actually needed in order to predict purchase probabilities has not yet been considered in research. Article [9] (see Section 3.2) gives insights concerning this research gap. The second aspect on which this dissertation focuses is the use of methods of social network analytics in combination with the prediction task. Social network analytics is "concerned with synthesizing the structural attributes of a social network and extracting intelligence from the relationships among the participating entities" (Gandomi and Haider, 2015, p. 142). This information can then be used to include additional information as features for the prediction task. These features are based on the underlying graph metrics of the resultant network structure. The combination of social network analytics to predict user behavior has been established through churn prediction in the telecommunication industry (Óskarsdóttir et al., 2017). Here, a relationship in terms of who contacts whom enables predictions to be made regarding churn behavior. For example, people are more likely to churn when they communicate often with another person who has already terminated their contract with the business. In this way, the relationships among individuals can be used to derive information from the resultant graph structure. However, the potential use of graph metrics for purchase prediction, as derived from a network structure based upon clickstream data, has not been extensively investigated in research. Thus far, only single graph measures (Byeon, 2013) or specific website-related metrics (Kalczynski et al., 2015) have been the subject of research. Other approaches have considered similarity between website visitors in order to detect community structures in social network graphs (Banerjee and Ghosh, 2001; Gündüz and Özsu, 2003). This dissertation therefore determines the potential of session-based user graphs derived from clickstream data as a means of predicting purchase probabilities in the e-commerce setting (see Section 3.2, article [8]).

2.3 Organizational Perspective: Social Media

The third dimension of information systems considered in this dissertation is the organizational perspective. Many businesses use information systems to support their entrepreneurial activity, regardless of the underlying industry. For some businesses this is especially relevant, since they are fully dependent on information systems and technological infrastructure. Businesses can be thus seen as high-level users of information systems. However, this thesis considers a related set of stakeholders, that is, the individual users and their behavior in regard to applications built upon the Internet infrastructure. In particular, two different areas of application form the focus. The first represents a special category of websites, i.e. social media platforms, while the second focuses on Internet access devices, and specifically how the usage of mobile devices in combination with ubiquitous mobile Internet access affects the everyday life of individuals in social and professional environments.

Social media can be defined as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and allow the creation and exchange of user generated content" (Kaplan and Haenlein, 2010, p. 61). At the heart of this definition is user-generated content, which is the content published by individuals who contribute to the social media website by providing and engaging with the available content. Therefore, the term social media comprises a broad concept of different websites based upon the Internet infrastructure, serving varying purposes. One way to classify them is based on the purpose of usage which is done by the social media landscape, which is a classification approach updated yearly by Fred Cavazza since 2008. The most recent landscape published in 2017 distinguished between six different main purposes of social media: collaboration, discussion, messaging, networking, publishing and sharing (Cavazza, 2017), with platforms and services combining several interwoven main functionalities. In general, social media websites can be used for entertainment, interaction with and observation of others, and information spreading and seeking, thereby providing a platform for both individuals and businesses to present themselves (Whiting and Williams, 2013).

Several more classification approaches exist which aim to characterize the vast number of existing social media websites into specific categories. Kaplan and Haenlein (2010) propose a two-dimensional classification approach. The first dimension these authors consider is the amount of self-presentation or self-disclosure prevalent on this website. Self-presentation or impression management involves the role a user plays in the online world by presenting their personal self with varying degrees of personal information (Trammell and Keshelashvili, 2005). The degree ranges from low to high, ordered according to the necessary amount of personal information with respect to whether the user or the content forms the focus. The second dimension considers the social presence of a user. Social presence involves the two dimensions of intimacy and immediacy (Kaplan and Haenlein, 2010). The former refers to the degree of personal contact between users, and the latter to whether the contact is delayed or takes place immediately. Here, three different levels are taken into account: low, medium and high. These two dimensions result in a 2x3 matrix consisting of six different categories of social media platforms: collaborative projects, content communities, blogs, social networking platforms, virtual games and virtual social worlds (see Table 1).

		Social presence and media richness		
		Low	**Medium**	**High**
Self-presentation / Self-disclosure	**Low**	Collaborative projects	Content communities [16]	Virtual game worlds [13]
	High	Blogs [11, 12, 15]	Social networking sites [14]	Virtual social worlds [13]

Table 1. Classification approach adopted from Kaplan and Haenlein (2010) for social media. The numbers in brackets indicate the relevant articles of this dissertation falling into each respective category (see Section 3).

The first category represents platforms with the lowest degree of self-presentation and social presence; these are typically seen as collaborative projects. Here the focus is not on the user but on mainly text-based content, produced in close collaboration with the community. Examples are wikis, where knowledge is gathered and presented in a structured way. Other categories where self-presentation is at a rather low level are content communities and virtual game worlds. Content communities are also centered around content distribution, but involve more personal content than collaborative projects. Examples of a virtual game worlds are massively multiplayer online role-playing games where an individual plays a fictional character unlike his or her real self; restrictions in terms of game play and settings available cause a low degree of self-presentation. However, recent games make it possible to create individual fictional characters, thereby causing a mixture of the different social media groups. Furthermore, media richness is very high, due to the complexity of game design (Kaplan and Haenlein, 2010).

Social media platforms where the user rather than the content is the focus are blogs, social networking sites and virtual social worlds. Blogs are a form of personal website where content is published in chronological order. In general, depending on the content published, two main types of blogs can be distinguished: personal and filter blogs. While personal blogs contain content with a high level of self-disclosure, filter blogs present and discuss external content (Trammell and Keshelashvili, 2005). Microblogs such as Twitter are a form of short, text-based blogs. Social networking sites are mainly used to connect with others, and the key aspect is the personal pages of individual users. Depending on the platform, varying types of media can be used. In the case of Facebook, which is the social networking site with the largest number of users worldwide (Statista, 2017d), both written and visual content in the form of photos and videos can be uploaded and shared within the user's personal circle. The last category, virtual social worlds, are a special form of virtual game worlds where individuals can create their own character and interact in a virtual world with other fictitious characters living a virtual life. The papers described in this literature review can be categorized using the classification approach of Kaplan and Haenlein (2010) depending on focus of the social media platform (see Table 1).

Social media applications depend on a rich base of users who provide the content and form the interactive environment. Operators of these websites therefore have an interest in providing their (potential) users with a satisfying user experience. To do this, it is necessary to understand who the users are and what they value most. Here, individualization is helpful, for example in providing users with the content they are most interested in. However, in addition to targeting individual users, the segmentation of users into specific groups is also possible. This segmentation of user groups can be carried out based on various characteristics. Possible

segmentation characteristics include the form and frequency of the user's interaction with the social media website (Lorenzo-Romero and Constantinides, 2012).
Additionally, user segmentation and individualization is not only of interest for social media operators who want to offer the user a satisfying experience, but also for businesses who can use these types of platforms as a medium for marketing activities to increase visibility and enable user interaction (Kietzman et al., 2011). Social media marketing is a specialization of online marketing, which also includes newsletter and online advertising campaigns and can be seen as "the use of social media channels to promote a company and its products" (Akar and Topçu, 2011, p.36). The use of social media as an online marketing tool constitutes both an opportunity and a risk for a company. Content created in form of posts and comments might create a positive or a negative reaction from users, and has the ability to be spread widely and rapidly through the social network through the sharing of content. Content on social media platforms produced by companies can be widely distributed by users sharing the content among their personal circle, and can therefore reach users that could not otherwise have been reached by the company alone. Furthermore, due to the possibility of direct and conjoint communication with users through social media profiles, there is the potential that the contact may be perceived as more personal, and the customer service of the business rated as more satisfactory (Gu and Ye, 2014). At the same time, reaction to content perceived as negative can also be amplified through the distribution of content throughout the network, causing a stronger visibility. Depending on the direction of this user-driven marketing, these have been referred to as either candy- or fire-storms. Marketing campaigns and communication with users through social media therefore needs to be carried out carefully. Here, a perfect fit between marketing campaigns and a specific user segment with certain behaviors and characteristics is valuable.
Since social media websites incorporate personal profiles, a segmentation based on demographics such as age, ethnicity and gender is easily applicable. In this regard, the main focus of this dissertation lies on the demographic variable of gender. Gender constitutes a social construct regarding how an individual perceives him- or herself and is one of the main characteristics by which individuals can be segmented into groups (Putrevu, 2001), since it is relatively simple to identify and generally comes in two different forms (Darley and Smith, 1995). Research has shown that gender has a strong influence on many areas of life such as how technology is used (Venkatesh and Morris, 2000). In the commercial context, gender has shown consistent differences in how and which products are bought, how information is processed and how users react to advertising (Cleveland et al., 2003).
As stated by Trauth (2013), research into gender in information systems is mainly separated into two areas. The first of these investigates the gender distribution among specific branches, and particularly analyzes the underrepresentation of females in certain industries. The second provides insights into how gender influences the adoption and usage of information technology. Research focusing on different social media platforms has reported different behaviors of males and females with regard to different aspects of usage. Gender theories aim to explain these differences and similarities from an evolutionary point of view.
This dissertation aims to offer insights into several aspects of social media research and user behavior. Firstly, research into gender differences in several social media categories is systematically reviewed in order to highlight differences between males and females regarding

usage and perceptions of social media applications in the areas of microblogs, blogs, online dating and online gaming platforms (articles [11-14]), to place existing research into a specific framework and to reveal potential research gaps. Furthermore, rather than a segmentation of users based upon gender, this dissertation also examines the special user group of German politicians and how they incorporate the microblogging platform Twitter as a tool during times of low political activity (article [15]). Here, methods of social network analytics (see Section 2.2) are applied to create a social graph representing the linkage structure among German politicians in order to understand how they interact with each other in terms of following and referring to colleagues.

Secondly, since social media platforms often make use of the personal information of users through the creation of profiles, the protection of highly sensitive user data is crucial. While platforms such as Facebook give users the possibility of restricting access to their personal profiles via specific privacy settings, content communities often refrain from personal profiles. Often, access to the available content on such communities is not restricted for users who are not registered but the interaction with other users or with the content (e.g. likes, comments) often requires an account. This is the policy used by the social media platform Reddit. The website claims that users can remain anonymous, since account creation requires a minimal amount of information in the form of a username and password. However, the privacy of users may still be in jeopardy through the content created by individual users. The possibility of de-anonymization of users through content is investigated in article [16].

Thirdly, handheld devices such as tablets and smartphones make mobile Internet access possible irrespective of time and location. In this regard, the propagation of smartphones around the world has enabled an all-time reachability of individuals and accessibility of information through the mobile Internet. Currently, 2.1 billion users around the world own a smartphone (Statista, 2017e). These devices are used regardless of location, situation and time, such as in the personal or professional environment, when alone or in the company of others (Do et al., 2011). The use of mobile phones as a side activity represents a form of multitasking, which is defined as engaging in more than one activity at the same time (Pashler, 1994). However, human beings are generally not able to parallelize, since they are prone to a cognitive bottleneck (Welford, 1967). The parallelization of several tasks leads to a decision-making process where fast switching between the different tasks simulates a form of multi-tasking, leading to a decrease in cognitive performance (Pashler et al., 2008). The ubiquitous nature of smartphones therefore leads to several situations where so-called multi-tasking is used. This influences the behavior of individuals regarding other activities and their environment. Research has shown that the mere presence of smartphones influences the perception of personal encounters in a negative way (Misra et al., 2014; Przybylski and Weinstein, 2012) and reduces task performance significantly (Thornton et al., 2014).

This dissertation focuses specifically on two different settings. In the personal environment, the influence of smartphone usage on romantic relationships is investigated from the perspective of jealousy (article [18]). In the professional environment, the effect of smartphone use during university lectures on the learning performance of students is analyzed, where performance is measured using the auditory and visual dimensions (article [17]).

3 Summaries of Articles

Overall, this thesis revolves around three main topics of research. Whereas the first one concentrates on the core of the thematic concept and focuses on the deeper understanding of the structure and the robustness of a specific network, i.e. the Internet topology, from varying vantage points, the second topic deals with the middle layer looking at different applications which are specific for and therefore being dependent on the Internet network. The third topic area concentrates on the systematic understanding of user behavior in the social media and e-commerce setting and is therefore directed at the outer layer of Figure 1. Here, the second main topic is subdivided into three chapters which are user behavior forecasts in e-commerce applications using predictive modeling methods, understanding gender differences on social media platforms via extensive literature surveys and understanding how the ubiquity of mobile Internet access due to smartphone devices has an impact on user behavior in different settings. In the following, summaries are provided regarding each paper submitted with this thesis.

3.1 Technological Perspective: Network Infrastructures and Applications

This chapter of this thesis deals with the analysis of the Internet topology to better understand its robustness in terms of random failures and targeted attacks. Each analysis is done focusing on another specific context such as investigating the robustness of the whole Internet graph or with respect to specific industries to understand the organizational structure of the Internet, including the geographical propagation of Internet components enabling a worldwide communication functionality.
A number of different applications are build upon the Internet infrastructure. One of the most important ones which is built upon the hypertext transfer protocol is the World Wide Web, forming the content of the Internet consisting of more than one billion websites connected with each other via hyperlinks. The World Wide Web consists of websites of different domains such as e-commerce and social media. Next to websites of the World Wide Web which will be in the focus of research in the following chapters, also other technological innovations have been built upon the Internet infrastructure. Especially two technological advances are in the focus of this dissertation which are the cryptocurrency Bitcoin and cloud service providers.

Article 1: *"How Robust is the Internet? – Insights from Graph Analysis"*

This paper contributes to the research question how robust the Internet network structure is towards targeted attacks and random failures. To do so, a dataset of the Internet network is collected to be able to construct it as a graph structure. On the level of autonomous systems, a set of four different data sources of the year 2012 is used and combined to derive a comprehensive map of the Internet consisting of around 44,400 nodes and almost 200,000 edges. Each node represents a specific autonomous system in the Internet network and they are connected via edges in case if there exists a possibility to exchange data between them. Using eight graph metrics commonly used in literature to characterize the structure of the network, the small-world nature is confirmed. Next the behavior of the network facing failures, i.e. the random deletion of nodes from the networks, and attacks, i.e. the targeted removal of nodes with a high degree, is investigated. Furthermore, the strategies of Xiao et al. (2008) are implemented focusing on connectivity-based attack modes. In the first step either a random or a specifically selected node are the starting point. In the next step, the failure or attack is spread through the network based on existing connections among those preliminary selected nodes. We show that the Internet network is highly robust towards the random outage of a high number of nodes. But due the importance of a few existing central connection points, targeted attacks do harm the network to a large extent and are therefore highly efficient. This shows a potential weak spot in the Internet infrastructure in case those high connectivity nodes are not well protected since potential alternative nodes are scarcely present in the Internet network.

Article 2: *"Vulnerability Against Internet Disruptions – A Graph-based Perspective"*

This paper extends the work from article one. The aim of the paper is to be able to classify nodes in the network, i.e. autonomous systems, according to their general risk towards failures and attacks based on their position in and relevance for the network. Using the same dataset as in case of article one, a connectivity risk score (CRS) is proposed which uses a combination of several graph metrics, unified into a single parameter which calculates a value for each node in the network. To be able to unify all chosen metrics into a single metric, they were normalized either using min-max normalization or an extended version of the z-normalization taking the special network setting into account. Finally, a specific weight is assigned to each metric, considering their relevance for robustness in the network. While low values of the CRS hint at a node being prone regarding random failures, high values do indicate that the node might be a favorable target for an deliberate attack. The analysis reveals that there are a lot of nodes which occupy a CRS value in the range between ten and thirty percent of the maximum value. Additionally, several nodes exist which have a very low value indicating autonomous systems being at the very outer border of the network. Very few nodes reach high CRS values, signifying that there are only some central points in the Internet network which are crucial for connectivity. All in all, based on the evaluation of the distribution of CRS values, most autonomous systems do have a potential for improvement regarding robustness.

Article 3: *"Towards Measuring the Geographic and Political Resilience of the Internet"*

Autonomous systems form the key component of the Internet and their physical locations are distributed all over the world. However, they are not spread homogenously and their number and density vary quite a lot depending on the country of origin. So far, existing papers did not consider the geographic location of autonomous systems when analyzing the robustness of the Internet network. Therefore, this paper proposes a new metric that measures the geographical Internet robustness on a country level. Even if direct interconnections are missing, the close proximity of autonomous systems within a short distance or on the country level hints at shared risks with respect to power outages or a central control through agencies. However, the assignment of autonomous system with respect to their exact location is error-prone. Therefore the focus of the paper lies on the country level to reduce potential mistakes. Combining several data sources, the assigned country for around 17,900 autonomous systems is identified. In the next step the location-based robustness metric is developed consisting of a combination of several measures based on country-specific characteristics such as population density and number of inhabitants in relation to the number of autonomous systems and IP addresses per country. Additionally, political indicators such as the world press freedom index are taken into account. Ranking the countries according to their combined geography-based robustness metric shows that mostly Oceania, North America and European countries can be found on top of the list whereas African countries are rather present at the end of the list. Asian and South American countries seem to be in a stage of transition towards a more developed Internet infrastructure.

Article 4: *"Who Runs the Internet? - Classifying Autonomous Systems into Industries"*

Autonomous systems form the Internet structure. They are owned and operated by specific technical administration entity. Understanding which entities form the Internet is of crucial interest since for example, some industries are more likely to be victims of cyber-attacks. Therefore understanding the heterogeneous character of the composition of the Internet can be helpful to increase its resilience. Using a publically available dataset of the Internet on the level of autonomous systems in combination with data originating from the U.S. Securities and Exchange Commission and regional routing registries, autonomous systems are classified into their belonging industry classes. The classification is done based on the North American Industry Classification System (NAICS). Using Standard Industrial Classification (SIC) codes, which can be mapped to the belonging NAICS class, in combination with a keyword-based approach around 57 percent of the 40,000 autonomous systems in the dataset could be assigned to at least one of the 18 industry classes. The most common industry classes are Internet service providers & networks, followed by trade & transport, education & research, telephone & communication and finance & insurance. Since not all autonomous systems could be automatically classified into industry classes further steps could use more sophisticated methods such as natural language processing to be able to derive a more complete picture of industries present in the Internet backbone.

Article 5: *“Exploring the Bitcoin Network”*

Bitcoins are an electronic payment method introduced in the year 2009. The cryptocurrency is based on cryptographic concepts and the blockchain where all transactions are stored. Up to this point several aspects of this technology remain still unclear. Taking an explorative look at the descriptive and structural components of the Bitcoin network, this paper heads into the direction to better understand the Bitcoin network, its usage and components.

For the analysis we use publically available blockchain data collected by the University of Illinois at Chicago (Brugere, 2013) in the time frame from 01/03/2009 to 04/10/2013. The dataset consists of 37.4 million edges and 6.3 million nodes. The analysis of the available data from a descriptive point of view shows that mainly very small transactions volumes of less than 10^{-5} Bitcoin take place. The average accumulated transaction value per day is around 910,000 Bitcoins with around 24,000 transaction per day on average. However, strong peaks in the Bitcoin exchange rate show up in times of high media coverage or collapsing financial systems.

In the next step a user-based graph is built based on the transactions made within the time frame where users represent the nodes in the graph and a connection exists between them if they made a transaction with each other. A year-based analysis to capture the dynamic character of the

transactions network shows a shift towards a scale-free network having several users who accumulated only a small number of transactions and few users who made a lot of them.
To demonstrate the possibility of de-anonymizing users in the transaction graph the authors select the largest node in the network and they are able to show that it belongs to the Bitcoin exchange Mt. Gox. In summary, the paper provides one of the first overviews on the descriptive and structural character of the Bitcoin network.

Article 6: *"Topological Analysis of Cloud Service Connectivity"*
When looking at the technical administration entity an autonomous system belongs to (see article [4]) one of the major industries is "IT and Internet services". Cloud service providers belong into this category, which is a service relying heavily on the Internet infrastructure. Disruptions in their service due to missing Internet connectivity can cause profit loss and damage the reputation of the service provider. Therefore understanding the connectivity risks of cloud service providers from the infrastructural point of view is necessary to see who is at risk to be able to provide potential countermeasures. So far existing research focuses on the improvement of data redundancy and connections within cloud data centers while this article analyzes the connection of cloud service providers beyond their own perimeter.
In the first step the authors conduct a survey of outages of the most important cloud service providers in the time period of 2008 to 2013 showing that disruptions caused by network failures are quite common causing around 17.5% of the total downtime for cloud service providers.
Based on a comprehensive map of the Internet topology, autonomous systems belonging to one of the most important cloud service providers worldwide are selected and their connectivity to the Internet network is analyzed using several graph metrics. Especially distance and centrality-based measures are relevant since they can be used to understand how well a node is connected to the rest of the network. We show that autonomous systems of cloud service providers are on average better connected compared to other autonomous systems in the network. According to our analysis, the cloud services of Google and Terramark seem to be the best connected so far. Further steps could analyze the connectivity risks of cloud service providers from a more fine-grained perspective using IP-level topology data to derive a more comprehensive picture how well those businesses are protected against outages from a infrastructural point of view.

3.2 Management Perspective: E-Commerce

The prediction of user behavior is in the focus of this part of the thesis. This subchapter includes papers submitted in the area of predictive modeling in the e-commerce context. E-commerce is one of the top applications built on top of the Internet infrastructure. Having a sales turnover of around $1.915 trillion and 1.6 billion online shoppers worldwide in 2016 (Statista, 2017f) this application has a tremendous effect on economy and society. Clickstream serves as a data basis of the analysis capturing the journey of website visitors in log files. Using a statistical model this behavioral information of website visitors can then be used to be able to predict how they will behave in the future. This enables companies to achieve a long-term viability and financial success due to being able to make more informed and therefore better decisions.

Article 7: *"Maximize What Matters: Predicting Customer Churn with Decision-centric Ensemble Selection"*

Understanding customer behavior and especially customer churn behavior is crucial for business success. The early detection of potential churn behavior might lead to the possibility of a timely reaction in form of countermeasures, preventing the customers' migration. Traditional performance measures do lack an intuitive interpretation helping to make informed management decision making. Therefore this paper uses the lift measure to derive a monetary-based measure of performance. Using a set of ten single classification models and a total of six ensemble learners the paper investigates whether the proposed decision-centric ensemble selection (DCES) method outperforms traditional approaches. The research questions are tested on eight real-life datasets and we show that the ensemble-based modeling approach increases profits in combination with retention campaigns by 0.47$ on average.

Article 8: *"Changing Perspectives: Using Graph Metrics to Predict Purchase Probabilities"*

The prediction of purchase probabilities in the e-commerce setting has been mostly done using features based on clickstream data. This paper tackles the problem from another perspective under the assumption that graph networks can be used to represent the user's journey on a website accurately enough to be able to predict their user behavior based on graph metrics.

The paper uses clickstream data from a two-month period of two e-commerce shops to derive user-based clickstream graphs. To do so, for each click of a user a new graph is created, updating each preceding graph of the user with new information. For each graph, i.e. for each click of a user, a selection of 23 different graph metrics is calculated. A multicollinearity analysis revealed that several metrics are highly correlated with each other resulting into a remaining set of 13 metrics. Those metrics are fed into three different classification algorithms, namely a generalized linear model and two non-linear models (gradient boosting machine and random forest model). Using area-under-the-precision-recall-curve and a lift chart as performance measures we see that the gradient boosting machine and the generalized linear model both outperform the random forest model on both datasets. Looking at the importance of each single feature we show that the graph metrics vitality, radius and the number of self-loops and circles in the user-graph seems to be promising in predicting purchase probabilities.

Therefore this paper adds to the existing research body another approach which might be promising for predicting user behavior in different contexts.

Article 9: *"The Price of Privacy: An Evaluation of the Economic Value of Collecting Clickstream Data"*

The collection of clickstream data happens whenever an Internet user visits a website on the Internet. This data is stored and used in varying ways such as marketing activities, therefore bringing a possible privacy hazard when collected without the explicit consent of the user. Deriving a set of 80 different clickstream features we classified them into four different categories each belonging to a specific level of privacy hazard. While the least privacy harming category contains features of the pages the user visited during the website journey, the second one adds behavioral aspects such as click-, scroll- and basket events the user performed while at the website. While the first two categories only use information of one user session at a time, the third one collects information over a longer time period spanning several user sessions. The last and most privacy harming category consists of possible identifiers such as location and user agent information. Using a dataset of two shops over a two month period and a non-linear random forest model we show that the least hazardous privacy setting already shows good purchase prediction results whereas the additional inclusion of the other settings adds no or only a marginal gain. Taking a look at variable importance we are able to confirm these insights since several features of the least privacy relevant settings are most important.

Article 10: *"Revenue Uplift Modeling"*
The estimation of the effectiveness of a marketing campaign is crucial to understand which marketing efforts are worthwhile and worth to allocate resources to. Established approaches fail to measure the effectiveness of a marketing campaign since those customers who performed the desired action due to the marketing action are not distinguished from those who would have done so nevertheless. Uplift models tackle this problem. Build upon A/B tests those models distinguish four groups of individuals. These groups consist of those individuals who received a marketing treatment and either responded or not and those who did not receive a marketing offer but either performed the desired action or not. The response of interest is then transformed in such a way that it represents this causality. This paper contributes to literature in that it provides a comprehensive overview on existing uplift methods and proposes a new approach which builds upon the maximization of revenue through an uplift model. For this revenue uplift model the transformed response variable is equal to one if the individual received a treatment and performed the desired actions resulting into a revenue exceeding a certain, pre-defined threshold. The response variable of individuals who performed the desired action but did not receive a treatment is set to zero which is also true for those individuals who did not perform the desired action at all no matter if they received a treatment or not.
Based on a huge real-world dataset of several e-commerce shops, the authors compare the predictive performance of a traditional model and established uplift modeling approaches with the new revenue uplift model proposed in the paper. Based upon the dataset applied, the main results of the paper show that the best performing uplift models do not outperform the best performing traditional approaches. However, the revenue uplift model is able to achieve the best results overall compared to all other approaches applied, achieving the largest increase in incremental revenue.

3.3 Organizational Perspective: User Behavior in Network Structures

This chapter of the thesis deals with the analysis of user behavior in the online context and offline context of Internet infrastructure applications. Especially the understanding of gender-related behavioral differences and similarities on social media platforms and the influence of smartphone technology in the professional and social context are in the focus. In the following, each individual subtopic will be explained in more detail.

3.3.1 User Behavior in Social Media

Social media websites can be used for self-representation, to stay informed about and in contact with others and to distribute content (Kietzmann et al., 2011). Facebook is one of the most important businesses worldwide in terms of market capitalization being in the top five of the Fortune 500 (Gandel, 2016). Having around 1.5 billion users resulting into a penetration of almost 23 percent of the world population (Internetworldstats, 2016), Facebook has a tremendous impact on business and society. While Facebook is mainly used to maintain relationships with others (Tosun, 2012), Twitter and other microblogging platforms main characteristic is seen as content and news distributor (Smith et al. 2012). Among such social networking websites, many other forms of social platforms used to engage with others and for entertainment do exist in the Internet such as blogs and online gaming.

This chapter mainly focuses on the literature-based analysis of gender differences among different types of social platforms present on the Internet. Additionally the de-anonymization of users and the usage structure of German politicians on social media websites is investigated.

Article 11: *"The Role of Gender in Blogging: Current State of Research and Future Directions"*

Blogs are a popular form of online communication enabling individuals to share their thoughts and knowledge. Understanding how the gender of the blog owner influences the perceptions of readers and the way the blog is operated is in the focus of this study. This is important since blogs have reached a tremendous recognition and readership count with 80 million unique visitors a month (Nielsen, 2012). Therefore their influence on society and economy is unprecedented. Using a meta-review in order to summarize existing research results the authors identify 48 articles focusing on gender-related insights in the blogosphere or report gender-based results as part of their study. The authors extract a total of 131 insights. Two coders independently classify those insights into the seven preliminary defined categories writing style, blog content, motivation to blog, bloggers' characteristics, privacy, readership and type of blog. Findings are mostly in line with Cross and Madson's (1997) self-construal theory which states that males are more independent striving for approval while females are more interdependent focusing on social aspects. This is also reflected in the insights related to motivation and types of blogs. While men prefer to write about professional topics to gain attention, reputation and to spread information, females prefer topics related to personal issues with a focus on forming social connections and to self-express. While gender is equally distributed among blog operators, males are perceived as more credible and do receive more attention and visibility. Contrary to the self-construal theory males do interact with other bloggers more often and support each other more frequently. The authors are able to identify the categories motivation to blog and privacy as the ones seldom addressed by research, therefore calling for more comprehensive research in those areas.

Article 12: *"Men, Women, Microblogging: Where Do We Stand?"*

With the rise of Twitter as the 140-character microblogging platform, these kind of short text-based communication platforms have become famous within the last couple of years. Having both millions of users, Twitter and Sina Weibo are two of the most used platforms around the world. The focus of this paper are on identifying gender-based differences and similarities with respect to the microblogging platforms mentioned above. Via an extensive literature survey using several scientific scholarly databases the authors are able to identify 60 studies published

between 2009 and 2014 which provide a total of 205 insights on gender-relevant findings. While 48 studies focus on Twitter, eleven use Sina Weibo and one paper uses both platforms. This shows a strong preference for the US-based platform. However, during their literature survey the authors only considered English-based studies.
The authors identify five leading themes which they name adoption, content, audience, motivation and presentation. Two coders independently classify each insight into one category. Most insights are evident for the themes presentation and audience with motivation lacking far behind all other categories, calling for more research in this area. Linking the insights with gender-based theoretical frameworks we show that several aspects of traditional gender roles are still evident on microblogging platforms. Females tend to share more among their personal circle and males are perceived as more credible. However, there is a shift observable that females begin to write more on male-seen topics showing that typical gender boundaries are crossed.

Article 13: *"Gender Differences in Online Gaming: A Literature Review"*

Online gaming is a phenomenon which has become even more popular with the rise of mobile devices, recording a rising number of users. However, insights on adoption, motivation, social interaction, self-presentation, performance and play patterns with respect to the gender divide remain unclear so far. Doing a literature survey based on several scholarly databases this paper sheds light on how males and females act similar and different with respect to online gaming. In total, the authors identify 47 papers providing insights on gender-based findings in the area of online gaming, classifying them into the six categories mentioned before. Findings are in line with traditional gender roles. While men are overrepresented in online gaming platforms, women are more likely to use these platforms to socialize with others. Also men do focus more on action game plays whereas women prefer logic-based games. However, with respect to motives to play similar patterns arise with games being a possibility to flee from reality and to have entertainment.

Article 14: *"Gender Differences in Online Dating: What Do We Know So Far? A Systematic Literature Review"*

Online dating platforms bring together females and males in anticipation to form relationships. Conducting a meta-review using a keyword-combination and a set of scholarly databases the authors find 69 relevant articles published between the years 1995 to 2015. Those research papers contain a total of 345 insights on gender-related findings with respect to the characteristics, the motivation, partner preferences, self-presentation, interaction and outcome of users using online dating platforms. Insights on partner preferences are by far the most common ones, whereas all others are distributed equally. The only exception are findings with respect to the motivation to use online dating platforms, where research cannot provide a lot of evidence so far. Results indicate that while males use and interact more on online dating platforms and are more likely to look for short-term relationships, females are rather motivated to use online dating platforms for long-term relationships such as in the romantic context or for friendship. Also for males physical characteristics are more important than for females, which on the other hand pay more attention to the socio-economic status of a potential partner. In terms of misrepresentation of personal characteristics to enhance their characteristics both gender make use of it. The results of this paper can help to develop better matching algorithms for online dating platforms. Further research in this direction can additionally focus on cultural differences.

Article 15: *"Twitter and the Political Landscape – A Graph Analysis of German Politicians"*
Twitter is one of the most important microblogging platforms having around 300 million monthly active user accounts in 2015 (Statista, 2017g). Among those one third have used the platform for political purposes such as to distribute political content and to encourage others to vote (Duggan and Smith, 2016). Therefore, politicians discovered the huge potential that lies within the use of microblogging platforms to reach potential voters and distribute their political views (Stieglitz et al., 2012). Among the 3,500 active politicians in Germany, around 1,700 of them operate a Twitter account. So far the political use of social media websites by German politicians has been mainly in the focus with elections nearby. Our paper aims at analyzing the interaction and connection between German politicians on Twitter via graph analysis in times where political actions are at a rather low frequency. This helps to understand the usage and interaction among German politicians on Twitter when there is no direct pressure to show presence to others. Relevant data of all Twitter accounts of German politicians is collected over a two-week period in August 2015 to construct three different graphs: the follower graph which represents the follower-followee relationship among German politicians and the mention graph which captures mentions of German politicians by their colleagues. The third graph is the intersection of both of them. While the CDU and the SPD are the parties with the most Twitter accounts, the Pirate party claims the highest number of followers. In general, Twitter is used frequently for interactions shown by several mention of peer politicians even during seasons of low political activity. Looking at the graphs and a set of metrics it becomes evident that the Green Party appears to be most connected while the AfD is the least connected and rather supports herself alone. CDU, SPD, Left Party and FDP appear to be quite homogenous. In general, German politicians seem to have a common approach on how to use Twitter.

Article 16: *"Privacy on Reddit? Towards Large-scale User Classification"*
Many social media platforms such as Facebook and Twitter are based upon the registration of users using their real identity. This can pose possible privacy threats since information and demographics stated on those websites can be used without the consent of users such as for marketing purposes or spam-related activities. Reddit is an online communication platform centered around anonymity. Upon registration only username and password are necessary to be able to participate actively on the platform. However, all content on the website is publically available. This paper contributes to the research question if it is possible to de-anonymize users based on their comment history. The characteristics social gender and citizenship are chosen as de-anonymization targets using an automatic approach. Using these characteristics it might be possible to de-anonymize users on this platform. In the first step, a dataset was derived from Reddit based upon 76,767 users who posted around 660,500 comments.
For the classification a supervised approach is selected which needs labeled data. Using regular expressions as suggested by Rao et al. (2010) and Pennacchiotti and Popescu (2011), users are

labelled based upon their comment history according to their stated gender and continent they live in. The authors compare the results of a weighted soft-margin support vector machine classifier (Osuna et al., 1997) and a supervised latent Dirichlet allocation (McAuliffe and Blei 2007) with features based upon the bag-of-words of the subreddits (i.e. the sub-forums of Reddit which are specialized on a specific topic, the user commented on). Using area-under-curve (AUC) as performance measure and a varying number of topics for LDA as parameter the authors show that in case of gender an AUC of 87.3% (two classes) and in case of citizenship and AUC of 53.8% (six classes) could be reached. This shows that it is possible to classify users on an anonymous platform according to social gender and citizenship solely based on the sub-forums the user is active on.

3.3.2 User Behavior and Mobile Devices

Smartphones are a wide-spread device of today's life, having around 2.3 billion users worldwide which is expected to be growing even further the following years (Statista, n.d.). With smartphones several applications such as e-commerce and social media have been made accessible from almost everywhere due to the mobile Internet. How the smartphone influences our daily lives – especially in the context of romantic relationships and in the academic context – is in the focus of this subchapter.

Article 17: *"To Phub or not to Phub: Understanding Off-Task Smartphone Usage and its Consequences in the Academic Environment"*

The presence of smartphones influences our daily life in many areas such as the in the professional (e.g., Roberts 2015) or personal environment (e.g., Karadag et al. 2015, McDaniel and Coyne 2016). One special professional area are lectures in universities where one professor talks to many students, creating an environment that enables students to use their device in between. The effects of such multi-tasking operations still remain unclear. Existing research investigates the effects of smartphone presence in lectures via single method approaches such as using surveys or experiments. Therefore, the authors strive to comprehensively understand the effects of smartphone use via a mixed-method approach. This approach includes observations, questionnaires, focus groups interviews and a quasi-experimental, naturalistic design (Babbie 2010; Silverman 2014). This is done to be able to understand how smartphone use during lectures influences the short-term performance of students which is measured using visual and auditory cues interspersed by the professor during the lecture.

In total, the article consists of three studies. The first study has the aim to observe the real smartphone use behavior of students during lectures connected with a subsequent survey to better understand the motivation of students to use their smartphones. The observations took place by two observers in a real lecture setting. In total, 60 students have been observed, of those more than 90% already had their smartphone present on the table. On average students used their smartphone eight times during the lecture. Most of the time the students used their smartphone for communication purposes or to browse the Internet. Self-reported motivations to use the smartphone are insufficient lecture style, boredom and the need to answer a message. The second study uses observations to be able to map self-reported with actual smartphone behavior. Additionally, a survey is distributed asking students to assess the current lecture, state

their self-reported smartphone use and to answer performance-related questions. Using a logistic regression the authors are able to show that the frequency of smartphone use negatively influences visual attention while the duration of smartphone use has a negative impact on the auditory attention of students.
The third study uses two focus groups to gain deeper knowledge why students use the smartphone during lectures. Main reasons to use are poor lecture style, low interest in the lecture and missing self-control. However, in the long run students do not see that the use of smartphones during lectures has an influence on their long-term performance. Possible countermeasures reported by participants of the focus group is to make the smartphone visual absent by leaving it in a place where it cannot be seen.
This study helps to derive a more comprehensive picture how smartphone use during lectures is performed and how the use of such devices might influence the short-term performance of students.

Article 18: *"Why Phubbing is Toxic for Your Relationship: Understanding the Role of Smartphone Jealousy Among "Generation Y""*

The permanent presence of smartphones can go together with positive as well as negative consequences. One possible negative influence has been shown by research with respect to romantic relationships (McDaniel and Coyne, 2016; Roberts and David, 2016). However, the drivers of the negative influence of partner's smartphone use behavior on relationships is not fully understood so far. To close this research gap this article focuses on jealousy as one possible driver. Jealousy can be seen as the loss of exclusive attention of the partner (Bauminger, 2010; Tov-Ruach, 1980) whereas the loss of attention can be caused by varying things such as other persons or objects as in case of the smartphone.
The authors conduct a two-part study with qualitative (study 1) and quantitative (study 2) elements. The first study's aim is to understand the emotions in relation to the smartphone use of the partner. To do so, the survey asked for the last time the partner used the smartphone for too long, the associated emotions and reaction. Most of the time the overuse of the device happened at home, before night's rest or during meals, whereas associated emotions ranged from positive to neutral to negative due to attention loss, sadness, anger and boredom. Common reactions are intervention, waiting behavior or simply no reaction at all. Some more drastic coping strategies such as to engage with other occupations and to express clear curiosity about the smartphone use of the partner were also present in the dataset.
Based upon the preliminary results of the first study, the second study focuses on understanding the role jealousy plays during the partner's smartphone use on their respective relational cohesion. Built upon a research model with the constructs partner's and personal smartphone use, jealousy and relational cohesion as well as several control variables such as age and gender the authors use the partial least squares (PLS) method to gain a deeper understanding of possible influencing factors. The authors show that the partner's smartphone use causes feelings of jealousy in the significant other whereas those feelings have a negative impact on the feeling of relational cohesion. Here, jealousy and the chosen control variables explain around one third of variance in relation cohesion, showing their tremendous impact on how ones romantic relationship is perceived when smartphone usage behavior is (omni-)present.

4 Publication of Academic Paper

In total, 18 articles have been submitted with this doctoral thesis. All of them are either submitted in conference proceedings, journals or are currently in the review process or working papers. Table 2 lists all papers submitted to proceedings and journals which appear in the three most relevant rankings for academic articles in the area of economics and information systems. The ranking systems chosen are the Handelsblatt Ranking BWL of the year 2012, the Erasmus Research Institute of Management Journals List ranking (EJL16) of the year 2016 and the ranking system of the Verband der Hochschullehrer für Betriebswirtschaft of the year 2015 (VHB-JOURQUAL3).

Of all articles, nine have been accepted to ranked proceedings and three of them to ranked journals. This accounts to around 67 percent of all papers submitted which have been accepted as a ranked publication according to the ranking systems chosen. Half of all articles of this dissertation are published in proceedings. Here, the most common proceeding is the *European Conference of Information Systems* (ECIS) with five articles published. According to the Handelsblatt Ranking BWL 2012 the ECIS proceeding is classified as 0.2 points (maximum is 1.0 points) and according to the VHB-JOURQUAL3 classified as B which is the third best option available in the ranking system. Other conferences with one paper each are the *Proceedings of the Hawaii International Conference on System Sciences* (HICSS), the *Proceedings of the Internationale Tagung Wirtschaftsinformatik* (WI) , the *Proceedings of the Americas Conference on Information Systems* (AMCIS) and the *International Conference on Information Systems* (ICIS).

In total, three articles submitted with this thesis have been accepted by peer-reviewed journals. Ranked with 0.4 points according to the Handelsblatt Ranking BWL 2012 and as B according to the VHB-JOURQUAL3 journal (third best option) the journal *Computers and Industrial Engineering* is one of the top published articles of this thesis. A second ranked journal article has been accepted by the *International Journal of Networking and Virtual Organisations*, ranked with 0.1 points according the Handelsblatt Ranking BWL 2012. Furthermore, one article has been accepted by the journal *Expert Systems with Applications* ranked as 0.2 according to the Handelsblatt ranking.

In addition to the papers mentioned above which have been submitted and accepted over the period of this doctoral thesis, one additional article is currently in the review process of a journal. This journal is *Business & Information Systems Engineering* (ranked as B according to VHB-JOURQUAL3). At the moment, the paper is in the second round of the re-submission process and is expected to be accepted for publication soon.

With respect to the order of authorship and based upon all papers submitted with the dissertation, in case of ten articles the author of the dissertation is on the first position, whereas half of those publications are ranked according to the three rating systems chosen. In case of five articles the author is on second position, with all publications being ranked. In case of two articles the author is on position three and in case of two articles on position four. This shows the substantial work the author of this dissertation has invested in the research in each of the articles submitted with this thesis.

	Rating system	Handelsblatt 2012	EJL16	VHB-JOURQUAL3	Impact Factor 2016	Articles published (articles submitted)
	Rating scale (from lowest to highest)	0.1 - 1	S, PA, P, P* / M*	E, D, C, B, A, A+		
Proceedings	Proceedings of the Americas Conference on Information Systems (AMCIS)	–	–	D	–	1
	Proceedings of the Hawaii International Conference on System Sciences (HICSS)	0.1	–	C	–	1
	Proceedings of the European Conference on Information Systems (ECIS)	0.2	–	B	–	5
	Proceedings of the Internationale Tagung Wirtschaftsinformatik (WI)	0.2	–	C	–	1
	Proceedings of the International Conference on Information Systems (ICIS)	0.2	–	A	–	1
Journals	International Journal of Networking and Virtual Organisations (IJNVO)	0.1	–	–	–	1
	Business & Information Systems Engineering (BISE)	–	S	B	3.392	1
	Expert Systems with Applications (ESwA)	0.2	–	–	3.928	1
	Computers and Industrial Engineering (CAIE)	0.4	S	B	2.623	1

Table 2. *Classification of articles submitted with this dissertation according to different ranking systems.*

References

Akar, E., & Topçu, B. (2011). An examination of the factors influencing consumers' attitudes toward social media marketing. Journal of Internet Commerce 10(1), 35-67. doi: 10.1080/15332861.2011.558456.

Albert, R., Jeong, H., & Barabási, A. L. (2000). Error and attack tolerance of complex networks. Nature 406(6794), 378-382. doi:10.1038/35019019.

Anderson, C., & Wolff, M. (2010). The Web is dead. Long live the Internet. Wired Magazine 18.

Babbie, E. (2010). The practice of social research (12th ed.). Belmont: Wadsworth.

Banerjee, A., & Ghosh, J. (2001). Clickstream clustering using weighted longest common subsequences. In: Proceedings of the Web Mining Workshop at the 1st SIAM Conference on Data Mining, 33-40.

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. Science 286(5439), 509-512. doi: 10.1126/science.286.5439.509.

Barabási, A.-L. (2016). Network Science. Cambridge University Press, 1st edition.

Baumann, A. (2013). Internet Resilience and Connectivity Risks for Online Businesses. Master thesis, Humboldt University of Berlin.

Bauminger, N. (2010). Jealousy in autism spectrum disorders (ASD). Handbook of Jealousy: Theory, Research, and Multidisciplinary Approaches, 267-292.

Belanger, F., Hiller, J. S., & Smith, W. J. (2002). Trustworthiness in electronic commerce: the role of privacy, security, and site attributes. The Journal of Strategic Information Systems 11(3), 245-270. doi: 10.1016/S0963-8687(02)00018-5

Bhattacharya, C. B. (1998). When customers are members: Customer retention in paid membership contexts. Journal of the Academy of Marketing Science 26(1), 31–44. doi: 10.1177/0092070398261004.

Brugere, I. (2013). Bitcoin Transaction Network Extraction. URL: https://github.com/ivan-brugere/Bitcoin-Transaction-Network-Extraction (visited on 07/18/2017).

Buckinx, W. & Van den Poel, D. (2005) Predicting online-purchasing behaviour. European Journal of Operational Research 166, 557-575. doi: 10.1016/j.ejor.2004.04.022

Byeon, H. (2013). Evaluating the online buying behavior using network analysis. International Journal of Advancements in Computing Technology 5(12).

Caida Ark (2013). Archipelago Measurement Infrastructure. URL: http://www.caida.org/projects/ark/ (visited on 07/17/2017).

Caida AS Rank (2012). AS Ranking. URL: http://as-rank.caida.org/(visited on 07/18/2017).

Cavazza, F. (2017). Social Media Landscape. URL: https://fredcavazza.net/2017/04/19/social-media-landscape-2017/ (visited on 07/18/2017).

Cebi, S. (2013). Determining importance degrees of website design parameters based on interactions and types of websites. Decision Support Systems 54(2), 1030-1043. doi: 10.1016/j.dss.2012.10.036.

Cleveland, M., Babin, B.J., Laroche, M., Ward, P. & Bergeron, J. (2003). Information search patterns for gift-purchases: a cross-national examination of gender differences. Journal of Consumer Behavior 3(1), 20-47. doi: 10.1002/cb.120.

Cohen, R., Erez, K., Ben-Avraham, D., & Havlin, S. (2000). Resilience of the Internet to random breakdowns. Physical Review Letters 85(21), 4626. doi: 10.1103/PhysRevLett.85.4626.

Cohen, R., Erez, K., Ben-Avraham, D., & Havlin, S. (2001). Breakdown of the Internet under intentional attack. Physical Review Letters 86(16), 3682. doi: 10.1103/PhysRevLett.86.3682.

Colgate, M. R., & Danaher, P. J. (2000). Implementing a customer relationship strategy: The asymmetric impact of poor versus excellent execution. Journal of the Academy of Marketing Science 28(3), 375–387. doi: 10.1177/0092070300283006.

Cross, S. E., & Madson, L. (1997). Models of the self: self-construals and gender. Psychological Bulletin 122(1), 5-37. doi: 10.1037/0033-2909.122.1.5.

Crucitti, P., Latora, V., Marchiori, M., & Rapisarda, A. (2003). Efficiency of scale-free networks: error and attack tolerance. Physica A: Statistical Mechanics and its Applications 320, 622-642. doi: 10.1016/S0378-4371(02)01545-5.

Crucitti, P., Latora, V., Marchiori, M., & Rapisarda, A. (2004). Error and attack tolerance of complex networks. Physica A: Statistical Mechanics and its Applications 340(1), 388-394. doi: 10.1016/j.physa.2004.04.031.

Darley, W. K., & Smith, R. E. (1995). Gender differences in information processing strategies: An empirical test of the selectivity model in advertising response. Journal of Advertising 24(1), 41-56. doi: 10.1080/00913367.1995.10673467.

Do, T. M. T., Blom, J., & Gatica-Perez, D. (2011). Smartphone usage in the wild: a large-scale analysis of applications and context. In: Proceedings of the 13th International Conference on Multimodal Interfaces (ACM), 353-360.

Dolev, D., Jamin, S., Mokryn, O. O., & Shavitt, Y. (2006). Internet resiliency to attacks and failures under BGP policy routing. Computer Networks 50(16), 3183-3196. doi: 10.1016/j.comnet.2005.11.010.

Dorogovtsev, S. N., & Mendes, J. F. (2013). Evolution of networks: From biological nets to the Internet and WWW. OUP Oxford.

Duggan, M., & Smith, A. (2016). The Political Environment on Social Media. URL: http://www.pewinternet.org/2016/10/25/the-political-environment-on-social-media/ (visited on 07/18/2017).

eMarketer (2016) Worldwide Retail Ecommerce Sales Will Reach $1.915 Trillion This Year. URL: https://www.emarketer.com/Article/Worldwide-Retail-Ecommerce-Sales-Will-Reach-1915-Trillion-This-Year/1014369 (visited on 07/18/2017).

ENISA (2017). Incident notification for DSPs in the context of the NIS Directive. URL: https://www.enisa.europa.eu/publications/incident-notification-for-dsps-in-the-context-of-the-nis-directive/at_download/fullReport (visited on 07/18/2017).

Erdös, P., & Rényi, A. (1959). On random graphs, I. Publicationes Mathematicae (Debrecen) 6, 290-297.

Faloutsos, M., Faloutsos, P., & Faloutsos, C. (1999, August). On power-law relationships of the internet topology. In: ACM SIGCOMM Computer Communication Review 29(4), 251-262.

Gandel, S. (2016). These Are the 10 Most Valuable Companies in the Fortune 500. URL: http://fortune.com/2016/02/04/most-valuable-companies-fortune-500-apple/ (visited on 07/18/2017).

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management 35(2), 137-144. doi: 10.1016/j.ijinfomgt.2014.10.007.

Gao, L. (2001). On inferring autonomous system relationships in the Internet. In: IEEE/ACM Transactions on Networking (ToN) 9(6), 733-745. doi: 10.1109/90.974527.

Garrie, D. B., & Wong, R. (2006). The future of consumer web data: a european/us perspective. International Journal of Law and Information Technology 15(2), 129-152. doi:10.1093/ijlit/eal022.

Gefen, D., & Straub, D. W. (2004). Consumer trust in B2C e-commerce and the importance of social presence: experiments in e-products and e-services. Omega 32(6), 407-424. doi: 10.1016/j.omega.2004.01.006.

Gu, B., & Ye, Q. (2014). First step in social media: Measuring the influence of online management responses on customer satisfaction. Production and Operations Management 23(4), 570-582. doi: 10.1111/poms.12043.

Gündüz, Ş., & Özsu, M. T. (2003). A web page prediction model based on click-stream tree representation of user behavior. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 535-540.

Hawkinson, J., & Bates, T. (1996). Guidelines for creation, selection, and registration of an Autonomous System (AS). Request for Comments. URL: https://tools.ietf.org/html/rfc1930 (visited on 07/18/2017).

Nielsen (2012). Buzz in the Blogosphere: Millions More Bloggers and Blog Readers. URL: http://www.nielsen.com/us/en/newswire/2012/buzz-in-the-blogosphere-millions-more-bloggersand-blog-readers.html (visited on 07/18/2017).

Internetworldstats (2016). Internet World Stats - Usage and Population Statistics. URL: http://www.internetworldstats.com/ (visited on 07/18/2017).

IRR.net (2013). Internet Routing Registry. URL: http://www.irr.net/ (visited on 07/18/2017).

Jarvenpaa, S. L., Tractinsky, N., & Vitale, M. (2000). Consumer trust in an Internet store. Information Technology and Management 1(1-2), 45. doi: 10.1023/A:1019104520776.

Kalczynski, P. J., Senecal, S., & Nantel, J. (2006). Predicting on-line task completion with clickstream complexity measures: A graph-based approach. International Journal of Electronic Commerce 10(3), 121-141. doi:10.2753/jec1086-4415100305

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. Business Horizons 53(1), 59-68. doi: 10.1016/j.bushor.2009.09.003.

Karadag, E., Tosuntas, S. B., Erzen, E., Duru, P., Bostan, N., Mizrak, S., Culha, I., & Babadag, B. (2015). Determinants of phubbing, which is the sum of many virtual addictions: A structural equation model. Journal of Behavioral Addictions 4(2), 60-74. doi: 10.1556/2006.4.2015.005.

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. Business Horizons 54(3), 241-251. doi: 10.1016/j.bushor.2011.01.005.

Kim, D. J., Ferrin, D. L., & Rao, H. R. (2009). Trust and satisfaction, two stepping stones for successful e-commerce relationships: A longitudinal exploration. Information Systems Research 20(2), 237-257. doi: 10.1287/isre.1080.0188.

Kopalle, P. K., Sun, Y., Neslin, S. A., Sun, B., & Swaminathan, V. (2012). The joint sales impact of frequency reward and customer tier components of loyalty programs. Marketing Science 31(2), 216–235. doi: 10.1287/mksc.1110.0687

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. Informatica 31, 249-268.

Kuan, H.-H., Bock, G. W., & Vathanophas, V. (2008). Comparing the effects of website quality on customer initial purchase and continued purchase at e-commerce websites. Behaviour & Information Technology 27(1), 3-16. doi: 10.1080/01449290600801959.

Kurant, M., Thiran, P., & Hagmann, P. (2007). Error and attack tolerance of layered complex networks. Physical Review E 76(2). doi: 10.1103/PhysRevE.76.026103.

Laudon, K. C., & Laudon, J. P. (2014). Management information systems: Managing the digital firm. New Jersey, 13th edition.

Liao, Z., & Cheung, M. T. (2001). Internet-based e-shopping and consumer attitudes: an empirical study. Information & Management 38(5), 299-306. doi: 10.1016/S0378-7206(01)00097-0.

Lorenzo-Romero, C., & Constantinides, E. (2012). Segmentation of users of social networking websites. Social Behavior and Personality: An International Journal 40(3), 401-415. doi: 10.2224/sbp.2012.40.3.401.

Magoni, D., & Pansiot, J. J. (2001). Analysis of the autonomous system network topology. ACM SIGCOMM Computer Communication Review 31(3), 26-37. doi: 10.1145/505659.505663.

Mahadevan, P., Krioukov, D., Fomenkov, M., Dimitropoulos, X., & Vahdat, A. (2006). The Internet AS-level topology: three data sources and one definitive metric. ACM SIGCOMM Computer Communication Review 36(1), 17-26. doi: 10.1145/1111322.1111328.

Mcauliffe, J. D., & Blei, D. M. (2008). Supervised topic models. In: Advances in Neural Information Processing Systems, 121-128.

McDaniel, B. T., & Coyne, S. M. (2016). “Technoference”: The interference of technology in couple relationships and implications for women’s personal and relational well-being. Psychology of Popular Media Culture 5(1), 85-98. doi: 10.1037/ppm0000065.

Metzger, M. J. (2004). Privacy, trust, and disclosure: Exploring barriers to electronic commerce. Journal of Computer-Mediated Communication 9(4). doi: 10.1111/j.1083-6101.2004.tb00292.x.

Misra, S., Cheng, L., Genevie, J., & Yuan, M. (2016). The iPhone effect: the quality of in-person social interactions in the presence of mobile devices. Environment and Behavior 48(2), 275-298. doi: 10.1177/0013916514539755.

Moe, W. (2003) Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. Journal of Consumer Psychology 13, 29-39. doi: 10.1207/153276603768344762

Moertini, V. S., & Ibrahim, N. (2015). Efficient techniques for predicting suppliers churn tendency in e-commerce based in website access data. Journal of Theoretical & Applied Information Technology 74.

Morisse, M. (2015). Cryptocurrencies and bitcoin: Charting the research landscape.

Nottorf, F., & Funk, B. (2013). The Economic Value Of Clickstream Data From An Advertiser's Perspective. In: Proceedings of the 21st European Conference on Information Systems, Utrecht.

O'connell, B. M., & Walker, K. R. (2014). U.S. Patent No. 8,650,080. Washington, DC: U.S. Patent and Trademark Office.

Oliveira, R., Pei, D., Willinger, W., Zhang, B., & Zhang, L. (2010). The (in) completeness of the observed internet AS-level structure. IEEE/ACM Transactions on Networking (ToN) 18(1), 109-122. doi: 10.1109/TNET.2009.2020798.

Óskarsdóttir, M., Bravo, C., Verbeke, W., Sarraute, C., Baesens, B., & Vanthienen, J. (2017). Social network analytics for churn prediction in telco: Model building, evaluation and network architecture. Expert Systems with Applications 85, 204-220. doi: 10.1016/j.eswa.2017.05.028.

Osuna, E., Freund, R. & Girosi, F. (1997). Support vector machines: Training and applications. Technical Report. Massachusetts Institute of Technology.

Padmanabhan, B., Zheng, Z., & Kimbrough, S. O. (2006). An empirical analysis of the value of complete information for eCRM models. MIS Quarterly, 247-267. doi: 10.2307/25148730.

Pai, D., Sharang, A., Yadagiri, M. M. & Agrawal, S. (2014). Modelling visit similarity using click-stream data: A supervised approach. In: International Conference on Web Information Systems Engineering, 135-145.

Park, S. T., Khrabrov, A., Pennock, D. M., Lawrence, S., Giles, C. L., & Ungar, L. H. (2003). Static and dynamic analysis of the Internet's susceptibility to faults and attacks. In: INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications, 3, 2144-2154.

Pashler, H. (1994). Dual-task interference in simple tasks: data and theory. Psychological Bulletin 16, 220–244. doi: 10.1037//0033-2909.116.2.220.

Pashler, H., Harris, C. R., & Nuechterlein, K. H. (2008). Does the central bottleneck encompass voluntary selection of hedonically based choices? Experimental Psychology 55(5), 313–321. doi: 10.1027/1618-3169.55.5.313.

Peck, J., & Childers, T. L. (2003). To have and to hold: The influence of haptic information on product judgments. Journal of Marketing 67(2), 35-48. doi: 10.1509/jmkg.67.2.35.18612.

Pennacchiotti, M., & Popescu, A. M. (2011). Democrats, republicans and starbucks afficionados: user classification in twitter. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 430-438.

Pitas, I. (2015). Graph-based social media analysis. Chapman and Hall/CRC.

Przybylski, A. K., & Weinstein, N. (2013). Can you connect with me now? How the presence of mobile communication technology influences face-to-face conversation quality. Journal of Social and Personal Relationships 30(3), 237-246. doi: 10.1177/0265407512453827.

Putrevu, S. (2001). Exploring the origins and information processing differences between men and women: Implications for advertisers. Academy of Marketing Science Review. URL: Available http://www.amsreview.org/articles/putrevu10-2001.pdf (visited on 07/18/2017).

Rachamin, O. (2014). How many online stores are there in the world? URL: https://www.digitalcommerce360.com/2014/12/04/how-many-online-stores-are-there-world/ (visited on 07/18/2017).

Radicati (2015). Email Statistics Report, 2015-2019. Available online at: http://www.radicati.com/wp/wp-content/uploads/2015/02/Email-Statistics-Report-2015-2019-Executive-Summary.pdf (visited on 07/18/2017).

Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010). Classifying latent user attributes in twitter. In: Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents (ACM), 37-44.

Roberts, J. A. (2015). Did you get phubbed by your boss?. URL: http://edition.cnn.com/2015/10/06/opinions/roberts-boss-phubbing/ (visited on 07/18/2017).

Roberts, J. A., & David, M. E. (2016). My life has become a major distraction from my cell phone: Partner phubbing and relationship satisfaction among romantic partners. Computers in Human Behavior 54, 134-141. doi: 10.1016/j.chb.2015.07.058.

RouteViews (2013). University of Oregon Route Views Project. URL: http://www.routeviews.org/ (visited on 07/18/2017).

Schneider, C. M., Moreira, A. A., Andrade, J. S., Havlin, S., & Herrmann, H. J. (2011). Mitigation of malicious attacks on networks. In: Proceedings of the National Academy of Sciences 108(10), 3838-3841.

Schweidel, D. A., Fader, P. S. & E. T. Bradlow (2008). Understanding service retention within and across cohorts using limited information. Journal of Marketing 72(1), 82–94. Doi: 10.1509/jmkg.72.1.82.

Shepherd, J. (2007). How to start a home-based online retail business. Globe Pequot, Guilford, Connecticut.

Silverman, D. (2014). Interpreting qualitative data. London: Sage Publications.

Smith, A. N., Fischer, E., & Yongjian, C. (2012). How does brand-related user-generated content differ across YouTube, Facebook, and Twitter? Journal of Interactive Marketing 26(2), 102-113. doi: 10.1016/j.intmar.2012.01.002.

Smith, P., Hutchison, D., Sterbenz, J. P., Schöller, M., Fessi, A., Karaliopoulos, M., Lac, C. & Plattner, B. (2011). Network resilience: a systematic approach. IEEE Communications Magazine 49(7), 88-97. doi: 10.1109/MCOM.2011.5936160.

Stange, M., & Funk, B. (2015). How much tracking is necessary? The learning curve in bayesian user journey analysis. In: Proceedings of the 23rd European Conference on Information Systems (ECIS), Münster, Germany.

Statista (2017a). Global digital population as of April 2017 (in millions). URL: https://www.statista.com/statistics/617136/digital-population-worldwide/ (visited on 07/18/2017).

Statista (2017b). Number of social media users worldwide from 2010 to 2020 (in billions). URL: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ (visited on 07/18/2017).

Statista (2017c). Number of digital shoppers in the U.S. 2019. URL: https://www.statista.com/statistics/251666/number-of-digital-buyers-worldwide/ (visited on 07/18/2017).

Statista (2017d). Most famous social network sites worldwide as of April 2017, ranked by number of active users (in millions). URL: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (visited on 07/18/2017).

Statista (2017e). Number of smartphone users worldwide from 2014 to 2020 (in billions). URL: https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/ (visited on 07/18/2017).

Statista (2017f). Number of digital buyers worldwide from 2014 to 2019 (in billions). URL: https://www.statista.com/statistics/251666/number-of-digital-buyers-worldwide/ (visited on 07/18/2017).

Statista (2017g). Ranking der größten sozialen Netzwerke und Messenger nach der Anzahl der monatlich aktiven Nutzer (MAU) im Januar 2017 (in Millionen). URL: https://de.statista.com/statistik/daten/studie/181086/umfrage/die-weltweit-groessten-social-networks-nach-anzahl-der-user/ (visited on 07/18/2017).

Sterbenz, J. P., Çetinkaya, E. K., Hameed, M. A., Jabbar, A., Qian, S., & Rohrer, J. P. (2013). Evaluation of network resilience, survivability, and disruption tolerance: analysis, topology generation, simulation, and experimentation. Telecommunication Systems 52(2), 705-736. doi: 10.1007/s11235-011-9573-6.

Sterbenz, J. P., Hutchison, D., Çetinkaya, E. K., Jabbar, A., Rohrer, J. P., Schöller, M., & Smith, P. (2010). Resilience and survivability in communication networks: Strategies, principles, and survey of disciplines. Computer Networks 54(8), 1245-1265. doi: 10.1016/j.comnet.2010.03.005.

Stieglitz, S., & Dang-Xuan, L. (2012). Political communication and influence through microblogging--An empirical analysis of sentiment in Twitter messages and retweet behavior. In: 45th Hawaii International Conference on System Science (HICSS), 3500-3509.

Tanenbaum, A. S. (2003). Computernetzwerke, Pearson Studium, 4th Edition.

Thornton, B., Faires, A., Robbins, M., & Rollins, E. (2014). The mere presence of a cell phone may be distracting. Social Psychology 45, 479-488. doi: 10.1027/1864-9335/a000216.

Tosun, L. P. (2012). Motives for Facebook use and expressing "true self" on the Internet. Computers in Human Behavior 28(4), 1510-1517. doi: 10.1016/j.chb.2012.03.018.

Tov-Ruach, L. (1980). When friends have other friends: Friendship jealousy in childhood and early adolescence. In: Handbook of Jealousy: Theory, Research, and Multidisciplinary Approaches. A. Oksenberg Rorty (Ed.). London: University of California Press, 465-488.

Trammell, K. D., & Keshelashvili, A. (2005). Examining the new influencers: A self-presentation study of A-list blogs. Journalism & Mass Communication Quarterly 82(4), 968-982. doi: 10.1177/107769900508200413.

Trauth, E. M. (2013). The role of theory in gender and information systems research. Information and Organization 23(4), 277-293. doi: 10.1016/j.infoandorg.2013.08.003.

Travers, J., & Milgram, S. (1967). The small world problem. Phychology Today 1, 61-67.

UCLA (2013). Internet Topology Collection. URL: http://irl.cs.ucla.edu/topology/ (visited on 07/18/2017).

Venkatesh, V., & Morris, M. G. (2000). Why don't men ever stop to ask for directions? Gender, social influence, and their role in technology acceptance and usage behavior. MIS Quarterly, 115-139. doi: 10.2307/3250981.

Welford, A. T. (1967). Single-channel operation in the brain. Acta Psychologica 27, 5–22. doi: 10.1016/0001-6918(67)90040-6.

Whiting, A., & Williams, D. (2013). Why people use social media: a uses and gratifications approach. Qualitative Market Research: An International Journal 16(4), 362-369. doi: 10.1108/QMR-06-2013-0041.

Woo, J. Y., Bae, S. M., & Park, S. C. (2005). Visualization method for customer targeting using customer map. Expert Systems with Applications 28(4), 763-772. doi: 10.1016/j.eswa.2004.12.041.

Wu, J., Zhang, Y., Mao, Z. M., & Shin, K. G. (2007). Internet routing resilience to failures: analysis and implications. In: Proceedings of the 2007 ACM CoNEXT Conference.

Xiao, S., Xiao, G., & Cheng, T. H. (2008). Tolerance of intentional attacks in complex communication networks. IEEE Communications Magazine 46(1). doi: 10.1109/MCOM.2008.4427244.

Yang, Y. C. (2010). Web user behavioral profiling for user identification. Decision Support Systems 49(3), 261-271. doi: 10.1016/j.dss.2010.03.001.

Zhang, Y., Bradlow, E., & Small, D. (2015). Predicting Customer Value Using Clumpiness: From RFM to RFMC. Marketing Science 34, 195-208. doi: 10.1287/mksc.2014.0873.

Part 2:

Submitted Articles

1 Technological Perspective: Network Infrastructures and Applications

ARTICLE 1:

HOW ROBUST IS THE INTERNET? – INSIGHTS FROM GRAPH ANALYSIS[1]

Authors: Annika Baumann, Humboldt-Universität zu Berlin, Berlin, Germany

Benjamin Fabian, Humboldt-Universität zu Berlin, Berlin, Germany

Proceedings: 9th International Conference on Risks and Security of Internet and Systems (CRiSIS 2014)

Publication year: 2014

Abstract

The importance of the Internet as today's communication and information medium cannot be underestimated. Reduced Internet reliability can lead to significant financial losses for businesses and economies. But how robust is the Internet with respect to failures, accidents, and malicious attacks? We will investigate this question from the perspective of graph analysis. First, we develop a graph model of the Internet at the level of Autonomous Systems based on empirical data. Then, a global assessment of Internet robustness is conducted with respect to several failure and attack modes. Our results indicate that even today the Internet could be very vulnerable to smart attack strategies.

[1] This article is provided with kind permission from Springer Nature. The original version is available at: https://doi.org/10.1007/978-3-319-17127-2_18.

1. Introduction

Cost efficient and fast worldwide communication is critically depending on the Internet; many innovative companies and services in the world rely on it to achieve their business goals. Therefore, reduced Internet reliability can lead to significant financial losses for businesses and economies. A study of an IT systems integrator estimated for example that network disruptions caused $1.7 billion financial losses for US companies in 2010 (CDW, 2011). But how robust is the Internet with respect to failures, accidents and malicious attacks? Internet failures already happened in reality, caused by natural disasters, power failures (Cerf, 2011), misconfiguration, or vicious attacks, and affected Internet connectivity (Agapi et al., 2011; Wu et al., 2007; Sterbenz et al., 2010).

Because of its size, structure, and dynamic nature, the Internet can be seen as a complex network. Several abstraction levels for representing the Internet in graph theory are possible: graphs of routers, points of presence, or Autonomous Systems (AS). An AS can be considered as an Internet domain, which is often under the control of a single organization such as an Internet service provider. Normally all entities assigned to one certain AS share a common routing policy, which is relevant for the traffic forwarding procedure. Because of rich data sources and a good research basis, we investigate Internet robustness at the AS level. Basing on empirical data we will develop a graph model of the Internet in order to conduct a global assessment of Internet robustness with the help of graph analysis.

2. Literature Review

Several graph models have been used to imitate the Internet structure. Classical network modeling approaches include the Erdös and Renyi (Erdös and Renyi, 1959; Newman, 2003) graph model (ER) and the scale-free BA model developed by Barabsi and Albert (Newman, 2003). Albert et al. (2000) study attack and failure tolerance of the Internet at the AS-level using both models. They focus on the diameter as a global connectivity metric. In Dolev et al. (2006), the Internet is studied at the AS level with restrictions caused by its economically driven character (policy-driven routing). Wu et al. (2007) examines resilience of the Internet in case of router removal. In Xiao et al. (2008), the authors develop attack techniques that are based only on local information. Schneider et al. (2011) also focuses on malicious attacks and develops a method for making the AS-level network more robust by interchanging edges of node pairs. In Sterbenz et al. (2010) and Smith et al. (2011) resilience frameworks, metrics, and case studies are presented. Deng et al. (2011) investigate the k-fault tolerance of the Internet.

3. Data Collection and Preparation

Collection and preparation of the dataset are crucial steps when studying the Internet topology. All publicly available sources suffer from incomplete or inaccurate data. We used three different approaches to gather data on the Internet structure: BGP routing tables, traceroute, and the Internet Routing Registry (IRR) during the same time period (mid of 2012). In the case of BGP routing tables it was possible to utilize data from other sources such as Oregon Route-Views (Route Views, 2014), RIPE-RIS (RIPE RIS, 2014), or from UCLA (UCLA, 2014). These projects integrate many data sources, namely BGP routing tables and updates, route servers as well as looking glasses. The Macroscopic Topology Project of CAIDA is based on Ark (CAIDA Ark, 2014) and collects IPv4 address paths with the help of several monitors located around the world, resulting in an AS link dataset. For IRR data, we retrieved data files of all available 34 IRRs (Internet Routing Registry, 2014). We only considered those AS paths that were changed not later than in 2012. Moreover, we demanded that any AS relationship is mentioned at least twice, by each of both participating ASs, for mutual verification. All of these individual datasets of

CAIDA AS Rank, UCLA, Ark, and IRR were merged into one final graph, resulting in a single connected component consisting of 44,397 nodes and 199,073 edges.

4. Internet Graph Statistics

Any graph can be characterized by various metrics, which are helpful for understanding its topological structure and connectivity. Table 1 presents an overview of selected relevant graph metrics and their respective values computed by our Python programs that are based on the NetworkX framework (NetworkX, 2014). An average clustering coefficient of around 0.46 indicates that almost half of all the possible connections between neighboring nodes are actually realized in the network. The AS-level graph is therefore quite well connected with respect to the neighborhood of each node; for comparisons to other networks see (Newman, 2003).

Metric	**Mean**	**Median**
Average Clustering Coefficient	0.4554	0.3333
Average Node Degree	8.9679	2.0000
Assortativity Coefficient	-0.1847	-
Average Eccentricity (Diameter / Radius)	7.8302 (11 / 6)	8.0000
Average Neighbor Connectivity	312.08	302.38
Average Neighbor Degree	703.29	315.00
Average Shortest Path Length	3.5585	3.5056
Average Node Betweenness Centrality	6.76296e-05	0.0000

Table 1. Graph metrics.

The average node degree of 8.97 indicates that on average nine edges are connected to a node in the AS-level network. The assortativity coefficient of the graph is -0.1847. This value implies that nodes in the network prefer to connect to other nodes having a dissimilar degree. In general, graphs with an assortativity coefficient below zero are considered more vulnerable to attacks because a large amount of nodes concentrates around those nodes with a high degree. The average neighbor degree with a value of 703.29 is yet another sign of the well-connectedness of the Internet. On average, every node in the network has 703 possible next hops to send its data to the desired destination. This is also confirmed by average neighbor connectivity (Mahadevan et al., 2006). Average eccentricity for the whole AS-level graph is around 7.83, indicating the average length of the longest shortest path from each node in the network. The Internet diameter at the AS level is 11. The average shortest path length is 3.56. This means that every possible pair of nodes can reach each other on average within less than four hops, which indicates a small world character (Watts and Strogatz, 1998). Average betweenness centrality calculates the number of shortest paths passing through a certain node, which roughly approximates the potential traffic flow through it. More than 24,000 nodes have a betweenness centrality of zero, i.e., there is no shortest path passing through that specific node. Nodes with a high degree usually also have large betweenness centrality because they are important for routing. These nodes usually form the center of the network and provide short routes through the entire network for other nodes.

To summarize, the Internet AS-level graph shows typical characteristics of so-called small world networks, in particular a high clustering coefficient as well as a small average shortest path length. Moreover, there are nodes with an extraordinarily high degree that represent central connection hubs.

5. Robustness Analysis

The robustness of the Internet topology can be analyzed by methods of graph theory. Our article will comprehend and update the findings of earlier robustness analyses based on more recent data. For this purpose, four different modes will be applied in order to examine the failure tolerance of the Internet. The first mode, random failure, is using the successive random deletion of nodes from the network. The second approach, the degree-based attack, involves the successive targeted deletion of nodes that are having the highest degree. This mode will simulate an organized attack that is disregarding any local links to select the next target node.

The third method, the mixed mode, is based on an approach proposed in Xiao et al. (2008). At first, a certain node will be chosen at random and removed from the network. Then, its former neighbors will be investigated whereupon the neighbor node with the highest degree will be deleted next. This approach will model a targeted attack that is restricted by limited knowledge because the necessary global information on the optimal next node might not be easily available with enough precision. This could simulate the continuous spreading of a worm or a similar threat through the network, starting at a random point and then trying to most efficiently distribute itself through the network based on local information. The fourth mode, the so-called random path mode, is similar to the third approach. Again, in the beginning a random node is chosen in the network and attacked. This time, however, the next neighbor node to be removed will be selected at random, moving continuously through the network using a randomized path. If a selected node becomes isolated, the next node for removal is chosen globally at random. This method simulates a realistic, randomized spread of malicious software or another threat through the network. This variant also utilizes the connections between nodes but not as efficiently as the third approach. For all modes, the results for the removal of the first ten percent of nodes will be presented. This is sufficient to show the impact of each mode.

To characterize the decay of the network in case of failures and attacks, we apply special metrics from graph theory that are describing connectivity of the network as a whole. One important global metric is the size of the giant component, i.e., the current size of the single largest connected component in the entire graph. Figure 2 shows the degeneration of this giant component in case of failures, attacks, the mixed, and the random-path modes.

The fastest network destruction is the degree-based attack. Only around two percent of the nodes need to be removed from the network in order to reduce the giant component to 50 percent of its original size. After eight percent of the nodes are eliminated, the former giant component has become negligibly small.

In case of the random failure mode, the size of the giant component decreases much slower, approximately by the amount of nodes removed from the network. Therefore, this mode has almost no effect on the network's global communication ability. The Internet is highly resistant in case of random and uncorrelated failures. In general, the mixed approach is a lot more efficient than the random elimination of nodes and for a particular amount of removed nodes even slightly more efficient than the attack mode. One reason could be that in case of the mixed approach the path-based selection of nodes with a high degree splits the network faster into a number of smaller components. Nevertheless, after a certain threshold is reached, the degree-based attack is consistently more destructive. In terms of destructiveness, the random path approach lies somewhere between these two highly efficient modes and the almost insignificant random deletion of nodes; the decay of the network is, however, much stronger than in the purely random case. The path-based selection criterion might again cause a fast splitting of components in the network topology. This illustrates the vulnerability of the Internet if a threat spreads throughout the network at random but is still aware of connecting paths. A lot of nodes with a low degree, whose removal does not affect the communication ability of the Internet by much, could still play an important role as some kind of glue holding the network together.

A similar result is provided by the metric of the total number of disjoint - but internally connected - components in the network (see Figure 1). In case of the random removal of nodes, only around 900 disjoint components arise after ten percent of nodes are eliminated. In case of the most efficient degree-based attack modus, the number of connected components grows rapidly. After ten percent of the nodes are missing from the network, the number of disjoint components is greater than 32,000. With mixed mode, after eliminating ten percent of the nodes almost 30,000 components exist - this is only 6.3 percent less than in previous case. The random path deletion of nodes again takes the third place in terms of efficient network destruction. Compared with the purely random deletion, it is again a lot more efficient due to its path awareness. Around 23,000 disjoint components arise after ten percent of nodes have been removed. Therefore path-based elimination of nodes is highly efficient in both cases, based on random or targeted neighbor selection. Furthermore, both approaches could be easier to perform since they are based on local properties rather than on global ones.

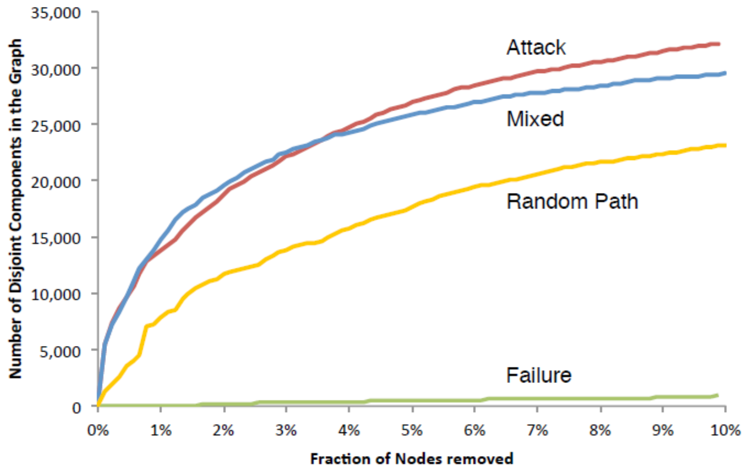

Figure 1. Number of disjoint components.

An even more detailed view on the network decay can be achieved by the distribution of the sizes of the connected components (Magoni, 2003). For this purpose, six distinct classes of component sizes are defined; each of them comprises components with certain amounts of nodes (Figure 2). Before the decay of the network starts, there is only one single connected component, which is part of class 5. To demonstrate the decay of the network in even more detail, the metric is refined in such a way that it measures the fraction of current nodes per connected component size class. This is the relation of the number of current nodes contained in each component class, divided by the total number of current nodes, i.e., all original nodes in the undamaged network minus the nodes already removed from the network. In case of the random deletion of nodes, there is almost no difference between the undamaged graph and the version having ten percent of its nodes lost. Class 5 dominates and its size is only reduced by the deletion of nodes and a very small number of components in classes 0 and 1. The decay of the network is therefore only driven by the deletion of nodes itself but not by further effects.

However, a different picture of the decay of the network is obtained when the nodes with the highest degree in the network are removed first. With such a degree-based attack, the fraction of nodes for component size class 5 drops rapidly. On the other hand, the fraction of nodes for class 0 rises fast continuously. After removing ten percent of the nodes, class 0 accounts for 73 percent of all remaining nodes in the network. This means there are almost 29,500 fully isolated nodes in the network. It is evident that the decay of the network in case of a targeted attack takes place much faster than in case of a random removal of nodes. The decay of the network in case of the mixed approach is quite similar to the degree-based attack modus. The random path approach does not generate such a drastic collapse as in the two previous cases. Again the curve of class 5 drops continuously but not as fast as in the attack or mixed mode. Even with partial randomness this is still a highly efficient mode to destroy a network.

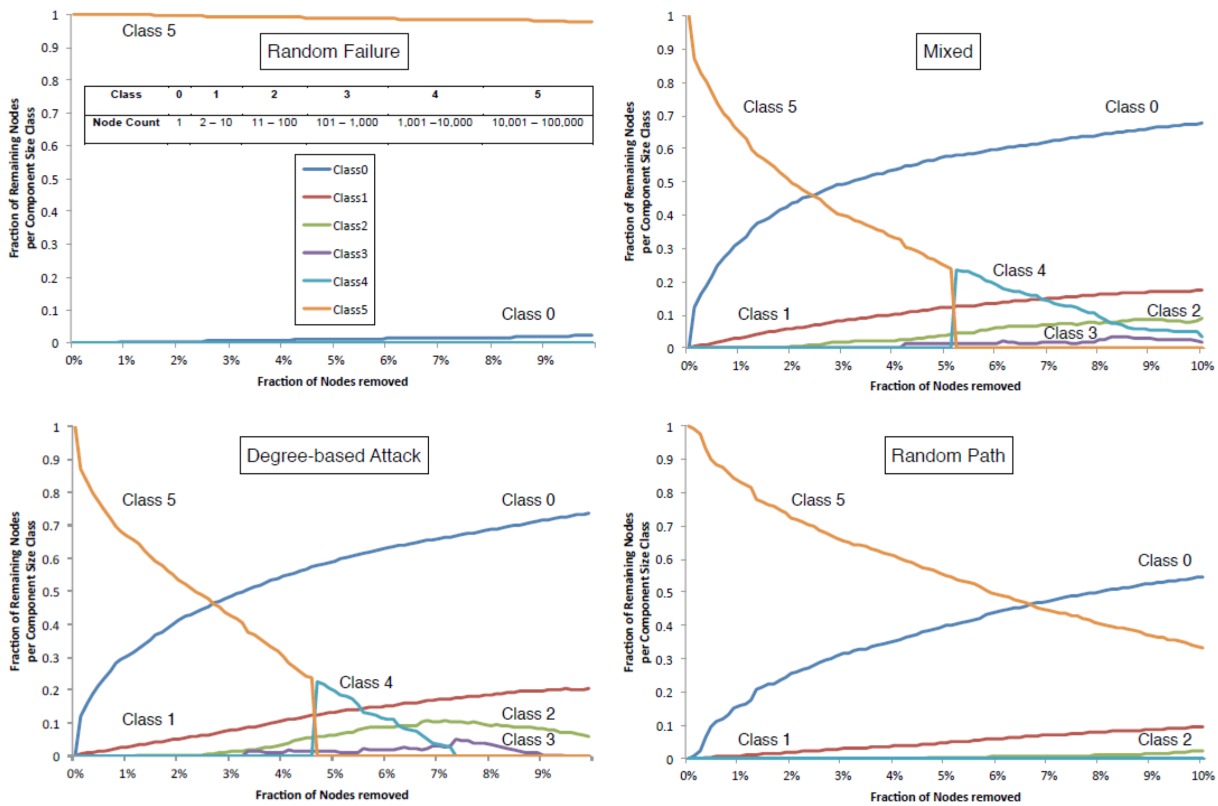

Figure 2. *Fraction of nodes per component size class.*

In summary, based on the total range of ten percent of node removal, the attack mode is the most efficient one. A disadvantage of this approach is the relatively high amount of information needed, which is based on the global graph and therefore not easy to obtain for any current Internet situation since most data sources are rather historical. If an initially faster but finally less destructive decay is adequate, the mixed approach is the better choice. But the major advantage of this mode lies in its strategy since no global information is necessary, which makes it easier to perform. By far not as efficient but still destructive is the random path method. This mode might be even easier to perform due to its randomness. It is sufficient to find any neighbor and then move along a randomly chosen path in the network.

6. Conclusion

Our results indicate that even today the Internet could be vulnerable to smart attack strategies, such as the degree-based attack, the mixed and even the random path attack mode. Our analysis is based on the abstraction level of ASs and on public data that is an incomplete snapshot. It is possible that valid links are not visible in our data, which would lead to somewhat better robustness results. In contrast, however, we would also like to further study the impact of economic relationships between ASs by considering policy-driven routing, and the very complex router level. Finally, we did not discuss practical IT security or fault tolerance of routing protocols. Our results motivate that further complementary studies are necessary, especially because the Internet has become a critical global infrastructure for businesses and everyday life.

References

Agapi, A., Birman, K., Broberg, R., Cotton, C., Kielmann, T., Millnert, M., Payne, R., Surton, R., van Renesse, R.: Routers for the Cloud: Can the Internet Achieve 5-Nines Availability? IEEE Internet Computing 15(5) (2011) 72-77.

Barabasi, A., Albert, R.: Emergence of Scaling in Random Networks. Science 286(5439) (1999) 509{512

CAIDA Ark. http://www.caida.org/projects/ark/ (2014).

CDW: Billions Lost due to IT Network Outages in 2010. http://www.eweek.com/c/a/Enterprise-Networking/Billions-Lost-Due-to-IT-Network-Outages-in-2010-Survey-559569 (2011).

Cerf, V.: Natural Disasters and Electric Infrastructure. IEEE Internet Computing 15(6) (2011) 103-104.

Deng, W., M., K., Mühlbauer, W., Zhu, P., Lu, X., Plattner, B.: k-Fault Tolerance of the Internet AS Graph. Computer Networks 55(10) (2011) 2492-2503.

Dolev, D., Jamin, S., Mokryn, O., Shavitt, Y.: Internet Resiliency to Attacks and Failures under BGP Policy Routing. Computer Networks 50(16) (2006) 3183-3196.

Erdös, P., Renyi, A.: On Random Graphs I. Publicationes Mathematicae (Debrecen) (1959) 290-297.

Internet Routing Registry. http://www.irr.net/ (2014).

Magoni, D.: Tearing Down the Internet. IEEE Journal on Selected Areas in Communications 21(6) (Aug 2003) 949-960.

Mahadevan, P., Krioukov, D., Fomenkov, M., Huffaker, B., Dimitropoulos, X., claffy, k., Vahdat, A.: The Internet AS-Level Topology: Three Data Sources and One Definitive Metric. ACM SIGCOMM Computer Communication Review (2006).

NetworkX. http://networkx.github.io/ (2014).

Newman, M.: Random Graphs as Models of Networks. In Bornholdt, S., Schuster, H.G., eds.: Handbook of Graphs and Networks. Wiley-VCH (2003).

Newman, M.: The Structure and Function of Complex Networks. SIAM Review 45(2) (2003) 167-256.

Outages-in-2010-Survey-559569/ (2011).

RIPE RIS. http://www.ripe.net/data-tools/stats/ris/ (2014).

Route Views. http://www.routeviews.org/ (2014).

Schneider, C., Moreira, A., Andrade Jr., J., Havlin, S., Herrmann, H.: Mitigation of Malicious Attacks on Networks. PNAS 108(10) (2011) 3838-3841.

Smith, P., Hutchison, D., Sterbenz, J., Schöller, M., Fessi, A., Karaliopoulos, M., Lac, C., Plattner, B.: Network Resilience: A Systematic Approach. IEEE Communications Magazine 49(7) (2011) 88-97.

Sterbenz, J., Hutchison, D., Cetinkaya, E., Jabbar, A., Rohrer, J., Schöller, M., Smith, P.: Resilience and Survivability in Communication Networks: Strategies, Principles, and Survey of Disciplines. Computer Networks 54(8) (2010) 1245-1265.

UCLA: Internet Topology Collection. http://irl.cs.ucla.edu/topology/ (2014).

Watts, D., Strogatz, S.: Collective Dynamics of 'Small-World' Networks. Nature (392) (1998) 440-442.

Wu, J., Zhang, Y., Morley Mao, Z., Shin, K.: Internet Routing Resilience to Failures: Analysis and Implications. In: Proceedings of the 2007 ACM CoNEXT Conference, New York, NY, USA. (2007).

Xiao, S., Xiao, G., Cheng, T.: Tolerance of Intentional Attacks in Complex Communication Networks. IEEE Communications Magazine 46(1) (2008) 146-152.

ARTICLE 2:

VULNERABILITY AGAINST INTERNET DISRUPTIONS – A GRAPH-BASED PERSPECTIVE[2]

Authors: Annika Baumann, Humboldt-Universität zu Berlin, Berlin, Germany

Benjamin Fabian, Humboldt-Universität zu Berlin, Berlin, Germany

Proceedings: 10th International Conference on Critical Information Infrastructures Security (CRITIS 2015)

Publication year: 2015

Abstract

The Internet of today permeates societies and markets as a critical infrastructure. Dramatic network incidents have already happened in his-tory with strong negative economic impacts. Therefore, assessing the vulnerability of Internet connections against failures, accidents and malicious at-tacks is an important field of high practical relevance. Based on a large integrated dataset describing the Internet as a complex graph, this paper develops a multi-dimensional Connectivity Risk Score that, to our knowledge, constitutes the first proposal for a topological connectivity-risk indicator of single Autonomous Systems, the organizational units of the Internet backbone. This score encompasses a variety of topological robustness metrics and can help risk managers to assess the vulnerability of their organizations even beyond network perimeters. Such analyses can be conducted in a user-friendly way with the help of CORIA, a newly developed software framework for connectivity risk analysis. Our approach can serve as an important element in an encompassing strategy to assess and improve companies' connectivity to the Internet.

Keywords: *Vulnerability, Internet Robustness, Internet Topology, Graph Mining, Risk Score*

[2] This article is provided with kind permission from Springer Nature. The original version is available at: https://doi.org/10.1007/978-3-319-33331-1_10.

1. Introduction

The importance of the Internet as today's communication and information medium is undisputed. It has revolutionized worldwide communication, made it cost efficient and fast and created countless of new or refined business models. There are numerous businesses in the world whose core competencies rely completely on the Internet. Based on these considerations, it becomes apparent that a limited or disrupted Internet connectivity can lead to significant financial losses for businesses and even economies. A study of the IT systems integrator CDW revealed that network disruptions caused $1.7 billion in financial losses in 2010 (CDW, 2010). This is an indicator of how crucial the Internet is for many business activities today.

In this paper, we aim to develop an analysis method and score that can help risk managers to assess the potential vulnerability of their organizations even beyond their own area of control, i.e., beyond their network perimeters. Here the question arises how robust their Internet connectivity is regarding failures, accidents and malicious attacks. How difficult is it to tear certain parts of their net-work neighborhood down? This article will examine this problem by first developing a global graph of the Internet based on a combination of several recent data sources. This graph will serve as a basis for robustness analyses focusing on the local vulnerability of single autonomous systems (ASs). From a high-level vantage point, an AS can be considered as an "atomic unit" of the Internet backbone, constituting a single administrative domain that is under the control of a particular organization, such as a company or public institution. Many companies in several industries own a dedicated AS (Baumann and Fabian, 2014), while for others the AS of their Inter-net service provider can be investigated.

From a topological point of view, connectivity risk can be characterized by being a potential victim of a random failure or a targeted attack. This leads to a certain duality: On the one hand, ASs that are not well-connected to the Internet are most at risk with respect to failures. On the other hand, those ASs, which are in-deed well-connected and therefore contribute most to communication ability and efficiency of the entire network, represent an attractive target for attacks aimed at weakening the global Internet. In order to make a statement about which ASs will fit into these two risk categories from a topological viewpoint, a proposal for a multi-dimensional score is developed in this paper that we call Connectivity Risk Score (CRS). This score is based upon a combination of selected and normalized topological metrics. Normalization ensures the comparability across ASs and network graph instances. Moreover, the CRS also reflects that the Internet topology is highly complex and the connectivity status of a certain AS depends on various factors.

The paper is structured as follows: First, related literature in the area of Inter-net resilience will be presented. Afterwards, the relevant methodology used in this paper will be described. Then, the development and evaluation of the CRS will be discussed. This score and all of the aggregated metrics can be accessed in a user-friendly way by security analysts via our newly developed CORIA analysis software that is presented in the subsequent section. The final section will summarize our contributions and results, discuss limitations as well as comment on future work.

2. Related Work

An important design feature of the Internet is its robustness. The term resilience can be seen as synonym and can be described as the ability of a certain system to return to a normal condition after external impacts. The robustness of the Internet is therefore "the ability of the network to provide and maintain an acceptable level of service in the face of various faults and challenges to normal operation" (Sterbenz et al., 2010, p.2). The approach presented in the current article extends established reliability analysis of online services, such as Tseng and Wu (2007) who focus on the reliability of critical servers, by analyses of connectivity based on the Internet graph.

Several researchers investigated the question of how to assess Internet robust-ness, but so far the main focus was placed on a global perspective (Baumann and Fabian, 2014). In an early work, Albert et al. (2000) analyzed the attack and failure tolerance of the Internet at the AS level based on both the classical Erdös–Rényi (ER) model and a scale-free graph model for the Internet. Dolev et al. (2006) additionally considered economical-ly driven restrictions of data exchange over the Internet backbone, i.e., policy-driven routing. Wu et al. (2007) examined the router robustness of the Internet in case of node removal, also taking policy restrictions into account. Xiao et al. (2008) focused on the attack tolerance of the Internet under the assumption that the possession of complete and global information about the Internet graph is an unrealistic assumption. Finally, Deng et al. (2011) considered the so-called k-fault tolerance of the Internet on the AS level which is the reachability of a pair of nodes in the network after the removal of k nodes.

More recent literature examines Internet resilience in a more specialized way. For example, Zhao et al. (2013) analyze the effect of removing the so-called k-core nodes from the Internet AS-level graph, i.e., the most important nodes which have at least degree k. Using a simulation based approach, Çetinkaya et al. (2013) propose a framework for better understanding the robustness of networks such as the Internet for future improvements. Shirazi et al. (2013) examine the resilience of anonymous communication networks such as Tor and propose a new metric for measuring robustness. Moreover, some projects already exist which examine the idea of combining different metrics to estimate the resilience of a network from a theoretical point of view (see for example Mieghem et al. (2010) and ResumeNet (2011)) where our approach will add significant results from a practical perspective.

3. Methodology

In order to obtain an extensive and recent dataset, we use a combination of three different main sources for Internet connectivity data: Border Gateway Protocol (BGP) routing tables, traceroute measurements and Internet Routing Registry (IRR) data.

In case of BGP routing tables, data provided by CAIDA's AS Rank project are used (CAIDA AS Rank, 2014) comprising a 5-day-period (06/01–06/05/2012). In addition, a research group of the University of California in Los Angeles (UCLA) provides another dataset (UCLA, 2014). Choosing the closest time period with available data files from 05/24/12 to 05/28/12, this dataset contains 159,383 unique AS paths. The traceroute-based Macroscopic Topology Project of CAIDA uses Archipelago (Ark) as a measurement tool (CAIDA Ark, 2014). All data files fitting into the appropriate time period (either the same as in case of CAIDA AS Rank or the most similar available) were downloaded from their website and preprocessed using only direct links between two ASs. After merging of the data, 57,922 unique AS paths are provided by Ark. In addition, Internet Routing Registry (IRR) data is also used in this paper. For this purpose, the data files of all available 34 IRRs were downloaded from their website (IRR.net, 2014). Based on the method mentioned by Siganos and Faloutsos (2005) as well as Zhang et al. (2005), the necessary AS path information was selected as a part of the aut-num object class. To gain reliable data only dyadic relationships were included in the dataset and those that were updated at last in 2012. The final IRR dataset consists of 47,348 unique AS paths.

All the individual datasets of CAIDA AS Rank, UCLA, Ark and IRR were then merged into one single file for the final dataset used in this paper resulting in 44,397 nodes and 199,073 edges.

4. Connectivity Risk Score (CRS)

4.1. Selection of Topological Metrics

To develop a comprehensive risk score, a literature survey was conducted examining existing metrics specifically used for assessing Internet robustness. Overall, 37 metrics could be identified[3]. Most of these metrics provide just a very general statement about the connectedness of an AS. Because of this, the CRS combines several metrics into a single measure to take advantage of multiple metrics and outweigh their disadvantages. Therefore in the next step, a number of requirements were defined which needed to be fulfilled by the metrics in order to be selected for the CRS. The initial properties that we required for the selection of metrics were:

1. A statement about the connectivity to the network for a single AS should be derived from it (not the entire graph, not aggregated AS groups).
2. The metric should have two distinct value ranges in order to distinguish the attractiveness for an attack and the susceptibility to failures.

Therefore, metrics that provide a statement solely for the global topology or AS groups were not selected (i.e., assortativity coefficient, symmetry ratio, (joint) degree distribution, average neighbor connectivity, eigenvalue-based metrics as well as global average metrics such as average degree, average clustering coefficient or diameter). Overall, six out of the initial 37 metrics remain which meet the requirements and are therefore used for the CRS: degree [DEG], average neighbor degree [AND], iterated average neighbor degree (two-hop neighborhood of a node) [IAND], betweenness centrality [BC], shortest path length [SPL] and eccentricity [ECC]. Some of these metrics are calculated based on the whole network structure (e.g., betweenness centrality) meaning that changing arbitrary nodes in the network might have an influence on the characteristics of that node. Those quasi-local metrics are still rather important for capturing the topological connectivity of a single AS since it cannot be seen as an isolated unit but is interconnected with a huge, interrelated network structure. For a more detailed description of these metrics see, e.g., Mahadevan et al. (2006) and Manzano et al. (2011). All of these metrics were calculated for the AS-level graph with the help of the graph analysis software NetworkX (2014) and average results for them are presented in Table 1.

	DEG	AND	IAND	BC	SPL	ECC
Average	8.9679	703.29	154.44	0.0001	3.5585	7.8302
Median	2.0000	315.00	95.573	0.0000	3.5056	8.0000
Max	4330.0	4330.0	4330.0	0.1300	7.8300	11.000
Min	1.0000	1.1400	1.1700	0.0000	2.1100	6.0000
Average Norm.	0.0018	0.1464	0.0350	0.0004	0.7470	0.6340
Median Norm.	0.0002	0.0641	0.0218	0.0000	0.7563	0.6000
Standard Deviation	60.385	901.43	202.03	0.0013	0.4425	0.5800

Table 1. Average results of metrics for AS data set (not normalized).

Correlation of the CRS candidate metrics based on the AS data set is shown in Table 2. None of the metrics are very highly correlated except for degree and betweenness centrality where there is still no perfect correlation, however. Based on these results, we conclude that the selection of metrics for the CRS is useful and non-redundant.

[3] The complete list of identified metrics is available from the authors upon request.

	DEG	AND	IAND	BC	SPL	ECC
Degree [DEG]	1.00					
Average Neighbor Degree [AND]	-0.05	1.00				
Iterated Average Neighbor Degree [IAND]	-0.07	-0.38	1.00			
Betweenness Centrality [BC]	0.85	-0.03	-0.03	1.00		
Shortest Path Length [SPL]	0.20	0.55	-0.33	0.10	1.00	
Eccentricity [ECC]	0.15	0.44	-0.25	0.07	0.74	1.00

Table 2. Correlation of normalized metrics on AS data set.

4.2. Normalization Process and Weighting of the Metrics

Because the results of the different metrics vary and also feature individual value ranges, they need to be normalized in order to calculate a composite score. For the metrics degree and betweenness centrality we used min-max normalization which maps the original range to the interval between [0,1]. In case of the metrics eccentricity and shortest path length – because here low values are desirable in terms of connectedness to the network – we used the max-min normalization.

For the (iterated) average neighbor degree we applied the z-normalization, which generates a normally distributed dataset with a new mean of zero and a variance of one due to possible distortions caused by high or low degree neighboring nodes. In order to consider the fact that nodes with only few neighbors tend to have a higher probability of having a median equal to the average neighbor degree, the number of neighbors was additionally taken into account. The resulting equation for the normalized (iterated) average neighbor degree of node i is as follows:

$$[I]AND_{corrected,i} = [I]AND_i + \left(\left(\frac{Median_i - [I]AND_i}{\sigma_i}\right) \cdot (\#of\ Neighbors_i)^{-1}\right) \cdot [\mathrm{I}]AND_i \qquad (1)$$

The up-voted and down-voted results for the (iterated) average neighbor degree were then finally normalized with the help of min-max normalization.

Furthermore, it is not reasonable to consider all metrics as equally important because their impact on the connection status of an AS might vary significantly. For the final CRS all metrics were weighted by a particular value. It was determined that the weights should sum up to one. The degree and the betweenness centrality were equally weighted and considered to be most important. Furthermore, the network environment one hop away from a node was considered to be more important for the robustness of a node than the network environment at subsequent hops due to the further distance from a node. Therefore, the average neighbor degree and the iterated neighbor degree were handled as less important and therefore weighted increasingly less. Because of the relatedness of the shortest path length and the eccentricity, the combination of both was weighted with the same amount as the degree and the betweenness centrality. Therefore, each of both distance-based metrics has a weighting of 0.125. This leads to the final weightings of 0.25 for the degree and the betweenness centrality, 0.125 for the shortest path length and the eccentricity, 0.15 for the average neighbor degree and 0.1 for the iterated average neighbor degree. The final score ranges between zero and one hundred percent.

4.3. Application of the CRS

Once the results for the selected metrics of the CRS are calculated, ranges need to be established that indicate critical areas for interpretation. The theoretical range of the CRS is between 100 % and 0 %, while the effective values, based on the dataset used, range between 74.58 % at the maximum and 0.01 % at the minimum. There are two numerical subranges that indicate those ASs that are most at risk. If the CRS is plotted against the risk of an AS in terms of robustness, a theoretical U-shaped curve is the result (see Figure 1, left), which can be regarded as exemplary representation used to visually communicate the idea of the CRS.

It is assumed that an AS is not at risk if it is neither vulnerable to random failures nor an attractive target for global attacks. This is associated with the following interpretation of the CRS: a small CRS value indicates a vulnerability to random errors. In the worst case, ASs that appear at the bottom of a CRS ranking list have a small degree and low (iterated) average neighbor degree. Their path options are quite limited. Furthermore, their betweenness centrality would in general be quite small, meaning that there are no or only few shortest paths passing through that node. Eccentricity and shortest path length can be expected to be high, indicating that these ASs are probably located somewhere at the edge of the network. Therefore, ASs with a low CRS value are badly connected to the network and insignificant for its communication ability, which makes them prone to random errors, but not attractive attack targets. On the other hand, a high CRS value is an indicator of high attractiveness for targeted attacks. In particular, ASs with a high degree as well as a high (iterated) average neighbor degree have a high CRS. Their path options are quite versatile. Both distance-based metrics are low. Those ASs are located at important communication points of the network, which can also often lead to a high betweenness centrality because many shortest paths are passing through those kinds of nodes. In summary, ASs with a high CRS value are well connected to the network and form an important communication backbone. This makes them highly attractive for deliberate attacks targeted at nodes whose removal would hurt the entire network most.

The distribution of the CRS values based on the empirical AS dataset is shown in Figure 1, right. Each data point shows a specific AS and its corresponding value of the CRS. There are many ASs with a low and few with a high CRS value. Only four of them reach a threshold of 50 % while two of them are extreme outliers having values of 74.58 % and 72.36 %. The average CRS value for all ASs is around 19.86 %. This again shows that there are many ASs having a low value, while the majority of them is located somewhere between 30 % and 10 %. This is an indicator that even today the global robustness of the Internet graph has a lot of potential for improvement.

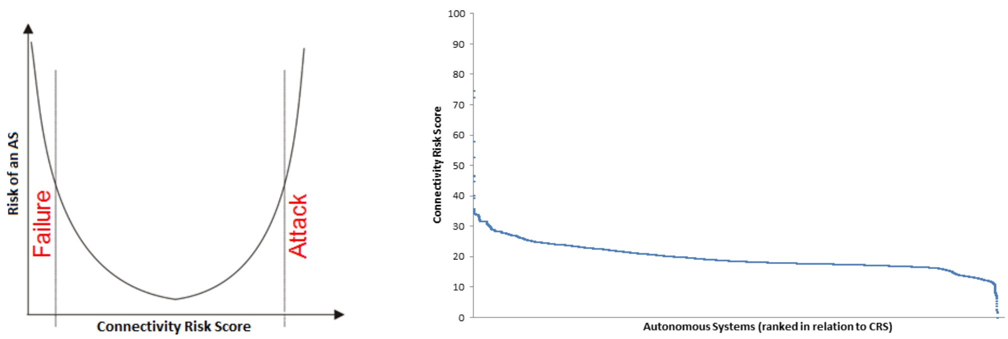

Figure 1. Critical areas for ASs (left) and distribution of CRS values (right).

A desirable global distribution of the CRS would involve a quite homogeneous accumulation of ASs in a certain score area to reduce the impact of attacks. This means that all ASs would be equally important for the communication ability of the network and the elimination of a selected AS would not have such a dramatic effect as it is the case now, e.g., more similar to the Erdös–Rényi model (Albert et al., 2000). At the same time, the CRS values should be as high as possible. A general increase of the CRS values for every ASs would be beneficial and would enhance the robustness of the global Internet graph by reducing the impact of random errors. We emphasize that the CRS serves as a first risk indicator to assess the vulnerability of single ASs but should be complemented in risk management practice by a more detailed examination of each particular AS.

4.4. Selective Validation of the CRS

In order to validate the usefulness of our CRS, we selected two poorly connected ASs with a low score and visualized their nearby graph environment (see Figure 2). Bold numbers represent the associated AS number, expressions in brackets specify the underlying organization; if an additional number is given, it refers to the degree of the nodes at the end of the network segment. The illustration makes it obvious that AS 636 and AS 45,076 are indeed badly connected to the Internet topology according to our dataset as was indicated by the CRS. For example, if any one of the subsequent one-degree nodes fail, these nodes will be affected as well and get completely disconnected from the rest of the network. Therefore, their connectivity depends not only on their own characteristics but also on those of the following ASs. In case of AS 45,076 the breakdown of any single node out of five (including the node itself) will affect this AS most severely. If only the degree metric were used, these ASs would not have scored worse than many other nodes with degree 1 and their particular vulnerability could have been easily overlooked.

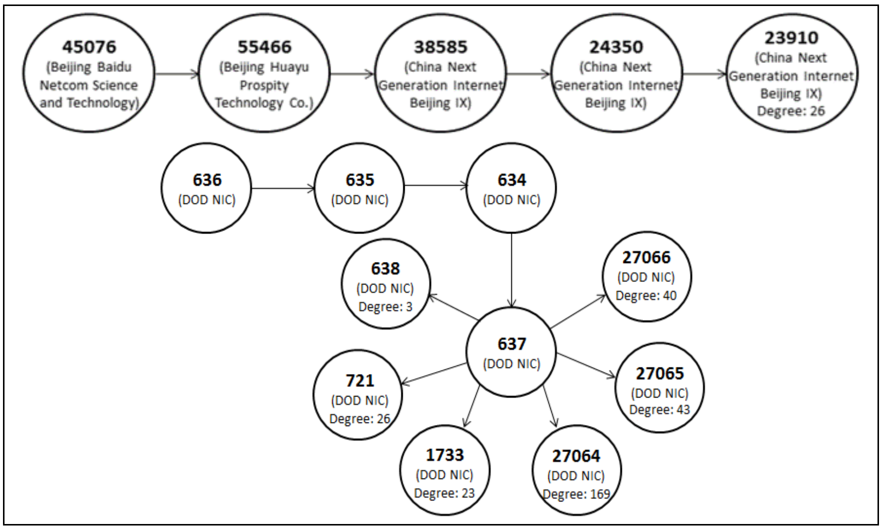

Figure 2. Two examples of badly connected ASs.

5. Connectivity Risk Analyzer (CORIA)

Based on the theoretical foundations presented in the earlier sections, we developed a web-based software framework that supports security analysts and consultants in assessing the connectivity risks of a particular organization. This *Connectivity Risk Analyzer* (CORIA) can be accessed via web browser. The analyst can search for individual AS and display the results of all connectivity metrics and the CRS discussed in this article. Moreover, statistical analyses of the entire AS datasets are possible and ASs can be ranked according to any of the metrics.

Metric	Duration [s]	Duration [m]	Percentage of Duration
DEG	0.468	0.01	0.000817%
AND	1.307	0.02	0.0022%
IAND	631.865	10.53	1.10%
BC	40226.493	670.44	70.26%
SPL	7553.916	125.9	13.19%
ECC	7456.873	124.28	13.02%
Total	57250.629	954.18	100%

Table 3. Calculation runtimes for the AS data set.

The high-level system architecture consists of a backend, the storage area as well as the frontend. The backend is responsible for importing AS data and calculating metrics and scores. It is implemented in the Python programming language and utilizes the network analysis framework NetworkX (2014). The calculation time of base metrics for the AS dataset used in this article are displayed in Table 3. All measurements were obtained in a virtual machine that was equipped with one CPU core running at 3 GHz and 4 GB of memory.

These results indicate that CORIA could cope with frequent updates of the underlying data set, which is one important direction for future improvements. The storage tier makes the results of the backend calculations persistent and enriches them with general information about each AS for later display via the frontend. It can also store interim results of calculations. For speed and flexibility, the storage tier is implemented using the Redis key-value store (Redis, 2014). User requests via a web browser are served by the frontend tier which is based on established web technology. It is implemented as a combination of a Ruby application based on the Sinatra framework (Sinatra, 2014) and HTML views based on the Twitter Bootstrap library (Twitter Bootstrap, 2014). An example user view is shown in Figure 3.

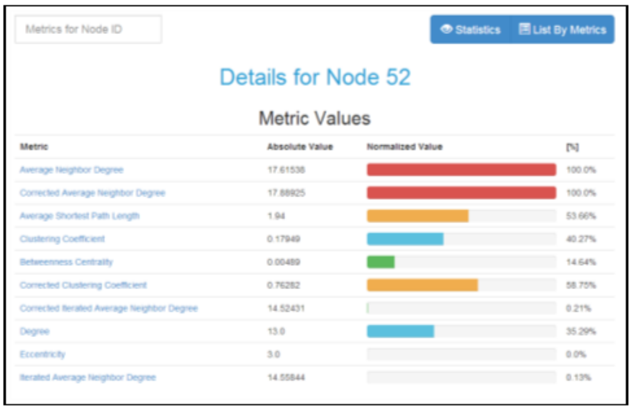

Figure 3. Example user view in CORIA (AS screen).

CORIA is designed with flexibility in mind: new or updated AS datasets can be loaded into the software whenever required. Further or refined metrics can be added with ease and can be flexibly combined into several different aggregated scores.

6. Summary, Limitations and Future Work

To the best of our knowledge, our proposed CRS marks the first attempt to measure the vulnerability of single ASs with regard to random failures and targeted attacks. The CRS takes several connectivity-based aspects into account and is therefore multi-dimensional (degree, average neighbor degree, iterated average neighbor degree, betweenness centrality, shortest path length, eccentricity). Considering these various dimensions helps to cope with the high complexity of the AS-level graph. In summary, the lower the CRS value, the more prone a specific AS is to random failures. The higher the CRS value, the more attractive a specific AS is for an attack. A desirable global state could involve a quite homogeneous accumulation of ASs in a certain high CRS value area.

Our article is subject to typical limitations of our research area that we aim to address in future work. Because there is a general lack of complete information regarding the Internet topology, our dataset is still incomplete. Also the dataset used in this work can be seen as only a first exemplarily starting point of investigation since it represents solely the AS level, is still incomplete in terms of included ASs as well as corresponding paths, and it originates from 2012. Furthermore, our data might contain some incorrect edges because its reliability strongly depends on the quality of the data sources used, e.g., the insertion of traceroute data whose alias resolution process is still not mature. Furthermore, policy-driven routing as well as traffic flow aspects are not considered so far but we plan to include this aspect in future work due to the possible current overestimation of viable connections in the Internet graph. Possible existing internal connections between ASs belonging to the same organization may not be visible in the public dataset. This might lead to imprecise risk assessment results based on the CRS only due to the underestimation of connectedness of the concerned ASs. However, an internal risk manager could take the CRS as a starting point of investigation and then verify the actual private connections of his or her organization.

In future work we also aim to further refine the CRS by developing an extended score which would take into account the homogeneity of ASs or rather their similarity of attractiveness in terms of being an attack target. In addition, we want to include an assessment of how hard it is to cut a certain AS off. It might also be valuable to utilize additional input for an extension of this score by using third-party knowledge. This could be achieved with the help of expert and stakeholder interviews.

Furthermore, statistical correlations between the current metrics and future metrics on different datasets should be investigated. Based on this, the weights could be adjusted accordingly in order to reduce internal correlations, improving the balance of the various metrics used in the score. Moreover, the score could then more reliably be transferred to other complex networks with different properties than the Internet AS graph. Therefore, we also plan to integrate further metrics. As research has shown, the effects of quasi-local metrics such as the spectrum of the graph, e.g., as in case of the spread of virus diffusions (Wang et al., 2003) which might be influenced by rather global metrics such as the assortativity coefficient (D'Agostino et al., 2012), show again the deep interconnection of the network where the general structure has an immense influence on characteristics of single nodes. Therefore in future research these effects of global metrics should not be underestimated and be taken into account.

We also aim to further validate the CRS and related newly developed scores. This could, for example, be conducted through several case studies such as in the application field of cloud computing (Fabian et al., 2015), an analysis of historical events of failure, a complementary IP-based analysis or insider knowledge of contracts between ISPs. Furthermore, we plan to enhance the web-based connectivity-risk analyzer CORIA with new features such as integrating further publicly available information on each AS and the possibility to manually add internal connections that may not be visible in the public data. Furthermore, each future evolution of the CRS and the underlying data set will be easily implemented in CORIA because it was designed for flexibility and extensibility. Another promising direction is adoption of methods from network reliability estimation, such as proposed by Lin and Chang (2011) for failures of cloud computing nodes, to our AS-level context. Not least, there is an important area of research on how to improve the connectivity of single organizations in a way that is technically and economically feasible. Since connectivity of an organization A does not only depend on the degree and local edges of A, one of the challenges involves how to motivate other organizations B and C to create links that primarily benefit A. Here, we aim to develop practical approaches based on game theory and economic mechanism design.

References

Albert, R., Jeong, H. and Barabási, A.-L. Error and Attack Tolerance of Complex Networks. Nature, 406, 378-382 (2000)

Baumann, A. and Fabian, B. How Robust is the Internet? – Insights from Graph Analysis. 9th International Conference on Risks and Security of Internet and Systems, Trento, Italy, Springer, LNCS 8924 (2014)

Baumann, A. and Fabian, B. Who Runs the Internet? Classifying Autonomous Systems into Industries. In: Proceedings of the 10th International Conference on Web Information Systems and Technologies (WEBIST), Barcelona, Spain (2014)

CAIDA Ark (2014). Archipelago Measurement Infrastructure. Retrieved April 30, 2015 from http://www.caida.org/projects/ark/

CAIDA AS Rank (2014). AS Ranking. Retrieved April 30, 2015 from http://as-rank.caida.org/

CDW. Billions Lost due to IT Network Outages in 2010: Survey. Retrieved April 30, 2015 from http://www.eweek.com/c/a/Enterprise-Networking/Billions-Lost-Due-to-IT-Network-Outages-in-2010-Survey-559569 (2011)

Çetinkaya, E. K., Broyles, D., Dandekar, A., Srinivasan, S., and Sterbenz, J. P. Modelling Communication Network Challenges for Future Internet Resilience, Survivability, and Disruption Tolerance: A simulation-based approach. Telecommunication Systems, 52(2), 751-766 (2013)

D'Agostino, G., Scala, A., Zlatić, V. and Caldarelli, G. Robustness and Assortativity for Diffusion-like Processes in Scale-Free Networks. Europhysics Letters, 97(6), 68006 (2012).

Deng, W., Karaliopoulos, M., Mühlbauer, W., Zhu, P., Lu, X. and Plattner, B. k-Fault Tolerance of the Internet AS Graph. Computer Networks, 55(10), 2492-2503 (2011)

Dolev, D., Jamin, S., Mokryn, O. and Shavitt, Y. Internet Resiliency to Attacks and Failures under BGP Policy Routing. Computer Networks, 50(16), 3183-3196 (2006)

Fabian, B., Baumann, A., Lackner, J.: Topological Analysis of Cloud Service Connectivity. Computers & Industrial Engineering, Volume 88, October 2015, 151–165.

IRR.net (2014). Internet Routing Registry. Retrieved April 30, 2015 from http://www.irr.net/

Lin, Y.-K., and Chang, P.-C. Maintenance Reliability Estimation for a Cloud Computing Network with Nodes Failure. Expert Systems with Applications, 38(11), 14185-14189 (2011)

Mahadevan, P., Krioukov, D., Fomenkov, M., Huffaker, B., Dimitropoulos, X., claffy, kc. and Vahdat, A. The Internet AS-Level Topology: Three Data Sources and One Definitive Metric. ACM SIGCOMM Computer Communication Review (CCR), 36(1), 17-26 (2006)

Manzano, M., Calle, E. and Harle, D. Quantitative and Qualitative Network Robustness Analysis under Different Multiple Failure Scenarios. In Proceedings of the 3rd International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (pp. 1-7) (2011)

NetworkX (2014). Retrieved April 30, 2015 from http://networkx.lanl.gov/

Redis (2014). Retrieved April 30, 2015 from http://redis.io/

ResumeNet (2011). Retrieved August 14, 2015 from http://www.resumenet.eu/

Shirazi, F., Diaz, C., Mullan, C., Wright, J., and Buchmann, J. Towards Measuring Resilience in Anonymous Communication Networks. In Proceedings of 6th Hot Topics in Privacy Enhancing Technologies (2013)

Siganos, G. and Faloutsos, M. Detection of BGP Routing Misbehavior against Cyber-Terrorism. In Proceedings of the 2005 IEEE Military Communications Conference (MILCOM 2005) (pp. 923-929) (2005)

Sinatra (2014). Retrieved April 30, 2015 from http://www.sinatrarb.com/intro.html

Sterbenz, J. P. G., Hutchison, D., Çetinkaya, E. K., Jabbar, A., Rohrer, J. P., Schöller, M. and Smith, P. Resilience and Survivability in Communication Networks: Strategies, Principles, and Survey of Disciplines. Computer Networks, 54(8), 1245-1265 (2010).

Tseng, J.C. and Wu, C.-H. An Expert System Approach to Improving Stability and Reliability of Web Service. Expert Systems with Applications, 33(2), 379–388 (2007)

Twitter Bootstrap (2014). Twitter Bootstrap Library. Retrieved April 30, 2015 from http://getbootstrap.com/about/

UCLA (2014). Retrieved April 30, 2015 from http://irl.cs.ucla.edu/

Van Mieghem, P., Doerr, C., Wang, H., Hernandez, J. M., Hutchison, D., Karaliopoulos, M., and Kooij, R. E. A Framework for Computing Topological Network Robustness. Delft University of Technology (2010).

Wang, Y., Chakrabarti, D., Wang, C., and Faloutsos, C. Epidemic Spreading in Real Networks: An eigenvalue viewpoint. In International Symposium on Reliable Distributed Systems, 25-34 (2003).

Wu, J., Zhang, Y., Morley Mao, Z. and Shin, K. G. Internet Routing Resilience to Failures: Analysis and Implications. In Proceedings of 2007 ACM CoNEXT Conference (CoNEXT '07), New York, NY, USA (2007)

Xiao, S., Xiao, G. and Cheng, T. H. Tolerance of Intentional Attacks in Complex Communication Networks. IEEE Communications Magazine, 46(1), 146-152 (2008)

Zhang, B., Liu, R., Massey, D. and Zhang, L. Collecting the Internet AS-Level Topology. ACM SIGCOMM Computer Communication Review, 35(1), 53-61 (2005)

Zhao, J., Wu, J., Chen, M., Fang, Z., and Xu, K. K-core-preferred Attack to the Internet: Is it more malicious than Degree Attack? In Proceedings of Web-Age Information Management (WAIM), Springer, Berlin Heidelberg (pp. 717-728) (2013)

ARTICLE 3:

TOWARDS MEASURING THE GEOGRAPHIC AND POLITICAL RESILIENCE OF THE INTERNET[4]

Authors: Annika Baumann, Humboldt-Universität zu Berlin, Berlin, Germany

Benjamin Fabian, Humboldt-Universität zu Berlin, Berlin, Germany

Journal: International Journal of Networking and Virtual Organisations 13(4)

Publication year: 2013

Abstract

Given the importance of the Internet for worldwide communication and services, its resilience against attacks, accidents, or attempts of misusing political control becomes critical for businesses and society. This article focuses on the question how vulnerable specific geographical regions are to an Internet access disruption or to censorship-based impediments due to governmental control. In particular, a new metric is developed that measures the geographical Internet resilience on a country level. For this purpose several indices based on geography, technology as well as control are combined into a single, rank-based score indicating the Internet resilience of a particular country compared to others.

Keywords*: Internet, Resilience, Autonomous System, Geography*

[4] This article is provided with kind permission from Inderscience Enterprises Ltd. The original version is available at: https://doi.org/10.1504/IJNVO.2013.064465.

1. Introduction

The importance of the Internet as today's communication and information medium is undisputed. To a large extent the current cost efficient and fast worldwide communication has been made possible by it. Moreover, it is the foundation for all online business models and services. The Internet is often defined as a global system of computer networks that uses the standard Internet protocol suite (TCP/IP) to serve users worldwide (Paul et al., 2013). In turn, a computer network, which is a collection of connected machines and IP routing prefixes under control of one or more operators with a common clearly defined routing policy, is referred to as an autonomous system (AS) (Hawkinson and Bates 1996). Today, the Internet consists of about 70,000 ASs (Potaroo, 2013) with a complex structure that complicates any kind of related analysis. Nonetheless, considering the role of the Internet in modern worldwide communication, social interaction, and economics it attracts significant scientific interest.

Internet resilience is a topic of highly practical political relevance which also has become a focus of several research studies from different disciplines. A limited Internet availability could lead to significant financial losses for economies and businesses. Furthermore, the ability to share information and communicate with people around the world increasingly depends on the Internet (ITU, 2011). Over the last decade, several Internet failures have occurred because of different reasons. Some were results of natural disasters, such as earthquakes (Cowie, 2011a), hurricanes (Brown, 2008a), and undersea mud volcano eruptions (Madory, 2013). Some were caused by a human factor such as cable cuts (Popescu, 2008b) or business disputes (Underwood, 2008c); and some originated from on-going battles for control over information streams on the web due to armed conflicts (Zmijewski, 2008d), politics (Cowie, 2011b), censorship (Brown, 2008e), central planning (Madory, 2012a), and terrorism (Madory, 2012b). All these examples highlight how numerous the Internet connectivity threats are, and that geography plays an important role in Internet resilience.

This paper contributes to answering the question how vulnerable specific geographical regions are to Internet access disruptions as well as censorship-based impediments due to local governmental control. For this purpose, we aim to develop a metric that measures the geographical Internet resilience based on a country level. This metric encompasses several geographic-, technology- as well as control-based indices, combining them into a single, rank-based score that estimates the Internet resilience of a particular country compared to others. Such a comparison could serve as an indicator for regions at risk, which could support international politics, Internet businesses or freedom activists in improving Internet resilience and censorship resistance by focusing their activities. The goals and methods proposed in this paper are motivated by investigating if and how an earlier approach by Roberts et al. (2011) could be improved by including additional metrics to account for further dimensions.

The structure of the paper is as follows: Section 2 discusses the background and current state of research on geographical vulnerability of the Internet. Section 3 describes the methodology and the methods used for data collection. Section 4 describes our new geographical resilience metric in detail. Finally, section 5 presents the results and the conclusions as well as limitations and future outlook.

2. Background and Related Work

Dramatic Internet failures already happened in reality. Sterbenz et al. (2010) as well as Wu et al. (2007) refer to various so-called "large-scale disasters". Even if the entire Internet as a whole seems to be quite robust (Sterbenz et al., 2010) these kinds of disruptions still affect the connection to the Internet in certain areas to a various extent. An example for a natural disaster is the hurricane Katrina, which raged in 2005 in the southeast of the USA – some of the most damaged areas were Louisiana (especially New Orleans), Mississippi, Alabama, Florida, and Georgia. A fraction of between 8 percent (in Louisiana) and 38 percent (in Mississippi) of the locally situated networks were unavailable due to Katrina. It was possible to restore most of the affected networks within a one-day period but especially networks in Louisiana needed a lot more time to recover fully (Cowie et al., 2005).

Another example for a natural disaster is the Taiwan earthquake in the end of 2006. The most severe impact was the damage of seven undersea cables in the affected area, when only two remaining cables still worked. Again, several networks were either unavailable or did not work properly – a maximum of around 4,000 networks were affected, most of them recovering quite fast after the seven earthquakes. The repair of the undersea cables took longer and lasted until the mid of February 2007 (Wilcox, 2007). The global impact of the Taiwan earthquake was stronger than the one of the hurricane Katrina because of the damage of the undersea cables. Locally, both natural disasters had a severe negative effect on Internet accessibility.

Other disasters, which are not necessarily of a natural origin, are for example fires caused by accidents. If such a fire occurs at a place important for the accessibility of the Internet, it might have a severe negative impact on it. In 2001 such an incident happened at the Howard Street Tunnel in Baltimore, Maryland. A fire caused by a train that ran off the rails destroyed fiber cables inside the tunnel. It was possible to reroute most of the traffic, resulting in slow connections for a bulk of Internet users. The repair of the damaged fiber cables required one and a half day (Sterbenz et al., 2010).

Next to these (natural) disasters a variety of further possible impediments exist which could affect the Internet access capability in a negative way. Mistakes – either on the human or system side – might have a severe impact too. A series of aligned faults caused a blackout in Canada and the Northeast USA in 2003. Not only around 50 million people had no access to electricity, but also the Internet access was significantly affected. A maximum of around 3,175 networks was unavailable due to the outage concerning over 50 percent of the ASs in this area and more than 1,700 organizations. For example, around 460 ASs had no working network left for more than four hours during the blackout – belonging to around 1,000 organizations (Renesys, 2003). Furthermore, it seems to be the case that organizations which are connected by many networks are only slightly affected. Organizations which have a small number of networks and are therefore only marginally connected suffer much more from the blackout, e.g., by having all their network connections lost. Therefore, the availability of the Internet depends not only on geography, its own infrastructure and functionality but also on several more factors like the power grid.

Other mistakes are accidental cable cuts such as the one of the Mediterranean Sea Cable in 2008, which seems to happen regularly (Sterbenz et al., 2010), and different forms of misconfigurations like the one by the Pakistan Telecom in 2008, which tried to block videos on YouTube by configuring their own AS as the path to the content. Instead of being valid only for Pakistan, this configuration spread worldwide. This mistake could only be finally undone by disconnecting the Pakistan Telecom. Another example for an accidental misconfiguration happened by a Czech Provider in 2009. A wrong configured BGP-setting of its router led to abnormally long AS paths affecting the worldwide Internet. An until then unobserved bug in Cisco's BGP implementation as well as a lot of misconfigured routers allowing these long AS paths led to this circumstance with global routing instability (Sterbenz et al. 2010).

Intentional attacks are another challenge including for example distributed denial of service attacks (DDoS), worms and acts of terrorism. The terrorist attacks on the World Trade Center on September 11 in 2001 caused severe damage not only on the human side. Important cables for Internet functionality in and next to the building were demolished. Although it affected around 1,000 networks, the effects on the global Internet were rarely noticeably. Most of the traffic could be rerouted or the necessary facilities could be rebuilt within a day (National Research Council, 2003). Furthermore, these kinds of events might even provoke another challenge to maintain the availability of the Internet which is unusually high traffic load caused by people seeking for information.

Worms are another form of intentional attacks targeting the Internet and its users directly. There are many examples of them such as the Nimda worm and Code Red v2, both raging in 2001. Apart from the destruction caused on private computers, these worms result in enormous BGP update rates leading to routing instabilities. In case of the Nimda worm, the update rates increased from around 400 per minute to around 10,000 per minute (Cowie et al., 2002).

In 2011, the European Network and Information Security Agency (ENISA) provided an Internet incident report. Eleven countries reported a total of 51 incidents, whereas nine countries reported no disturbances and another nine countries just sent in no report at all. The most frequent incidents were caused by hardware/software as well as third party failures – accounting for 80 percent of the errors. Human errors, malicious attacks and natural impacts are less frequently the cause of the disturbance. Nevertheless, natural disasters such as storms and earthquakes as well as malicious attacks have the most severe impact causing a much longer duration of the disruption of service than the other three impact categories. This incident report makes obvious that there are many sources of interference causing more than 50 significant Internet service disruptions. However, the estimated number of unreported or undetected cases might be much higher (ENISA, 2012).

All of these examples hint at the importance of considering geography as an important factor of Internet resilience. Several research papers already contributed to the question of how robust certain countries are, focusing on the AS level of the Internet topology. For example, Roberts et al. (2011) used data collected by CAIDA (CAIDA AS Names, 2012) and Team Cymru (Cymru, 2013) to perform an AS-to-country mapping. They used the results of the mapping to construct a metric reflecting the complexity of the Internet infrastructure in a single country, which incorporates the number of IP addresses, the number of ASs, and the number of points of control, which are defined as ASs which provide connection to 90% of IP addresses in a country. Their findings indicate that absolute size does not always determine the complexity and resilience of a national network infrastructure. While there are 177 ASs in China, Roberts et al. identified that there are only 4 points of control. Thus, a threat of a country-wide Internet outage is much higher compared, for instance, to Hungary where the number of ASs and the number of points of control are 143 and 17, respectively.

Reynolds and Tamaddon (2011) extended the results of Roberts et al. (2011) with Internet filtering scores. They used several rankings provided by the OpenNet Initiative to construct two entities – the Political Filtering Score and the Overall Filtering Score. The combined metric is called the NetworkPolitical Resiliency (NPR) and indicates the internal power of a country to control its own Internet access (Reynolds and Tamaddon, 2011). In accordance with their calculations, only one country out of 152 analyzed was classified as High NPR – the United Kingdom, while the majority, 101 countries, received a Low NCR rating.

Wählisch et al. (2012) proposed an enhanced IP-block-based approach for identification and classification of ASs related to a specific country. They argue that foreign ASs hosting national organizations should also be considered as a part of national network infrastructure. Thus, to identify such ASs, Wählisch et al. (2012) suggested shifting from the prefix-based level to the IP-block level. The researchers employed the developed technique to spot all of the Germany-related ASs and claim to outperform the prefix-based methodology by 25%.

All of these papers considered solely technology-based metrics to measure the vulnerability of the Internet on the AS-level with respect to certain countries. We will extend this approach by including additional metrics to assess further dimensions and therefore enhance the analysis to a different level.

3. Methodology and Data Collection

In this section, we provide our methodology and describe the data collection. First, we will discuss the concept of geographic proximity in the Internet. Besides the topological vulnerability of ASs, there is another type of vulnerability which is independent of existing connections. This is the vulnerability due to geographic location. One variant of this vulnerability is characterized by a very close geographic proximity of ASs, e.g., if the power supply to a particular town fails in which several ASs are located. In this case, several ASs might be affected at the same time even if there is no actual topological connection between them. Thus, investigating the close geographic proximity of ASs is challenging.

Therefore, a methodology of geo-locating ASs will be provided. However, depending on the degree of accuracy and detail it might not even be possible to specify for some ASs whether they are close to each other. Thus at first, the geographic location of IP addresses needs to be analyzed. Services such as the IxMapper of Ixia (IxMapper, 2013) or EdgeScape of Akamai (EdgeScape, 2013) mainly use hostname-based mapping methods to find a solution for this (Lakhina et al., 2003). The degree of correctness of GeoIP services is questionable. MaxMind (Maxmind, 2013a), a supplier for GeoIP Services, provides on its website accuracy results for IP geolocations based on an allocation to cities. Correct results within a distance of 40 kilometers range between 45 and 96 percent, while the average value is 73 %. This means that on average 27% of the allocation results are either wrong or the city is unknown (MaxMind, 2013b). In the second step, these IP addresses need to be assigned to the corresponding ASs. This so-called alias resolution process is not yet fully developed and can cause errors during the assignment procedure. This kind of more detailed geographic analysis is conducted, for example, by Lakhina et al. (2003). But due to the immaturity of this method for the close geographical proximity of ASs, it will not be considered in this paper.

The second type of geographic vulnerability is characterized by the fact that ASs are situated in the same country, i.e., their wide geographic proximity within country borders. In such a case it might happen that a whole country is cut off from the Internet. This could happen, for example, if it is easy to control the ASs in a certain area. This might be caused by the fact that there are only few ASs regulated by a small number of responsible companies. Recently such Internet cut-offs happened in Syria and Egypt (Cowie, 2011c; Cowie, 2012c). This analysis is not as strict as the close geographic proximity analysis. The allocation within country borders covers a larger geographical area than a city does, which makes it possible that mapping errors are reduced. This paper will therefore focus on the geographic analysis of ASs situated in the same country.

In addition to the IP geolocation method mentioned above, another approach for retrieving the geographic location of an AS is viable. The Regional Internet Registries (RIRs) contain in their databases a two-letter country code based on ISO 3166 for each AS that indicates the country in which the managerial unit is located or rather where the AS has been registered (APNIC Aut-Num, 2013; Roberts et al., 2011 p.4). This may be the same as the geographic location of the AS but this does not necessarily have to be the case. As mentioned by Roberts et al. (2011, p.4), the country code seems to be accurate enough for such an analysis. There may be some exceptions such as ASs in Africa that have been registered somewhere else in the world due to the earlier nonexistence of the AfriNIC.

The geographic resilience analysis conducted in this paper is therefore based on the data source provided by Team Cymru (Cymru, 2013). This is a service offering listings that obtain daily updated information about country codes directly from the various RIRs including ARIN, RIPE, AfriNIC, APNIC and LACNIC. The data was downloaded on February 18th, 2013. All information on ASNs was selected from these datasets and finally merged. In total there are 54,773 ASNs mentioned in combination with their underlying country code. For the final analysis, entries containing too general location information such as ‘Europe’ or ‘Asia/Pacific region’ were ignored.

To validate the reliability and dynamics of this data, it was cross-checked with the country code data provided by CAIDA’s AS Rank project (CAIDA AS Names, 2012) and the GeoIP data from MaxMind (MaxMind, 2013a). The AS Rank country data (CAIDA AS Names, 2012) was used to investigate how likely it is that country codes (frequently) change. The data is based on the same method, i.e., it provides country codes collected via various RIRs. The difference lies in their date of creation. The AS Rank data originates from the 29th of June 2012 resulting in difference of 235 days between these two datasets. By comparing the intersection of ASs of both data files, it was possible to find out how many country codes have changed. Both data files share 52,306 joint ASNs. Of these, 2,176 ASNs changed their underlying country code, which is around 4.16 percent. In general, the country data remains quite stable over time. A closer look at the data reveals that around 250 entries are specifications of general country information, e.g., general country codes, such as ‘EU’ (Europe) or ‘AP’ (Asia/Pacific region), are stated more precisely in the Cymru data and changed, for example, to ‘HU’, ‘AT’, or ‘JP’. In such a case, this should not be regarded as a real change but as an increase of precision. This means that in summary around 3.7 percent of the entries effectively changed. The dynamics of country codes is therefore present but negligible and thus considered to be quite stable.

In order to analyze the reliability of the Cymru dataset and its country codes, the MaxMind database is used. The MaxMind website provides GeoLite Country (IPv4 and IPv6) as well as a linked GeoLite ASN (IPv4 and IPv6) databases. This is a GeoIP service which might also not be 100 percent reliable but could still be seen as more accurate than country codes of the RIRs. The databases are updated on the first Tuesday of every month. Therefore, they originate from the 5th of February 2013. This results in a difference of 13 days between the Cymru and MaxMind data. All databases for IPv4 and IPv6 were merged together to create a list of all ASNs with their underlying GeoIP location. If more than one country was mentioned for one and the same ASN, all entries related to that ASN were discarded. This was done in order to achieve unambiguous results. Remaining entries that contain a country code such as “satellite provider” or “anonymous proxy” were also dismissed. To ascertain the level of data reliability, the following procedure was applied: at first, the intersection of joint ASNs in both the Cymru and the MaxMind datasets was identified. Overall they share 17,857 common ASNs. In the next step, the intersection of ASNs was checked for changes in terms of the underlying country. The country changes for 814 ASNs, i.e., a fraction of 4.55 percent of the ASs. This difference is not excessively severe. Even if the country code is not highly reliable, the majority of information seems to be correct. In conclusion, the country code information derived from Cymru is reliable enough to conduct a deeper analysis without possible errors caused by GeoIP data.

4. Geographic Resilience Ranking of ASs

The approach proposed in this paper is based on a similar method designed by Roberts et al. (2011). This paper will update some of their results as well as include more metrics that are reflecting different dimensions of Internet robustness. We propose to include the following metrics into the geographic resilience score for specific countries:

1. Absolute number of ASs per country;
2. Number of ASs per a square kilometer per country;
3. The ratio of the number of ASs to the number of inhabitants per country;
4. Number of ASs in relation to the population density of a country;
5. Absolute number of IP addresses per country;
6. Ratio of the number of ASs per number of IP address per country;
7. Number of IP addresses per capita per country;
8. Risk score of becoming a target for cyber-attacks;
9. World Press Freedom index.

In the following sections we will describe the motivation for choosing each metric as well as the data sources used to gather the necessary information. Only those countries and their underlying coding were considered as relevant which are listed in the ISO 3166 coding list on Wikipedia (Wikipedia ISO 3166 Coding List, 2013). Of all 268 countries mentioned in the coding list of Wikipedia, 231 are also mentioned in the data provided by Team Cymru. This leads to 37 missing countries which are either not contained in the dataset or do not have an AS on their own. For the analysis, a rank is assigned to each country for each used metric resulting in nine distinct rankings for each country. This positions the country according to the value of the particular metric. For example, the country with the highest number of ASs will be ranked on position one while the country with the lowest number of ASs will have the position number 231. In the end, all assigned ranks for each country will be summed up and basing on the sum of ranks a new final rank will be calculated. This will give an indication of the risk for a country to be disconnected from the Internet. The rank-based approach was chosen due to its simplicity and easy practical application.

4.1. Geography-based Metrics

Our proposed resilience score combines several geography-based metrics which are the absolute number of ASs per country, the number of ASs per square kilometer per country, the ratio of the number of ASs to the number of inhabitants per country and the number of ASs in relation to the population density of a country. For this purpose, the geographic area in square kilometers, the population and the population density for every country was determined via Wikipedia.

The first metric is the only absolute one. The total number of ASs indicates how complicated it might be to cut a country off the Internet. A higher absolute number of ASs means that there are more entities which provide Internet infrastructure. It is an important but not sufficient metric in terms of robustness. In general, this metric favors countries that span a large area. They usually have a higher number of ASs present, while small countries and islands possess only few ones due to their small size. Therefore, the absolute number of ASs per country might be misleading. A small country such as the Vatican having an area of 0.44 square kilometers will never have such a large number of ASs as Italy spanning an overall area of 301,338 square kilometers.

Hence, a ratio is applied which takes into account the area a country embraces. It calculates the number of ASs that are present per square kilometer of land – the area ratio. It considers the size of a country and thus the area which needs to be covered by the number of ASs present in that country. It is assumed in this approach that in general the bigger a country is, the more ASs are usually needed to provide reliable Internet access. This time small countries have an advantage. Even if they have just one AS due to their small size, their ratio is usually higher than in case of large countries.

This results in the circumstance that not just the area of a country is important. It is necessary to address countries with a large area but a small population size in a better way. For example, Greenland is a country that has only a small number of present ASs and spans at the same time a large area. Therefore, it is disadvantaged in both metrics mentioned above. Considering its low population density, there even seems to be no need for a better provision of ASs. To address those countries more efficiently, two additional metrics are applied. A ratio related to the population gives an indication on how many people have to share an AS (population ratio). If the available number of ASs cannot serve the number of inhabitants, this could, for example, lead significantly faster to traffic overloads, which in turn might cause a disturbance of the networks. Therefore, this ratio needs to be in balance, too.

Taking into account only the metric based on the population size of a country might be misleading because it also favors small countries. This kind of countries usually have a small number of inhabitants due to space limitations and therefore often a favorable population ratio. Thus, in addition the quotient of the number of ASs and the population density is used (density ratio). Again, this metric refers to the geographical size but takes the existing population size into account at the same time. Therefore, it calculates the area that is covered by an AS per inhabitant. Densely populated countries have a higher demand in terms of AS availability. While a solely population-based metric produces an advantage for small countries, this is not true for the population density. Here, big countries with a low population density have an advantage. However, this is only the case if they simultaneously have a sufficient number of ASs in their geographical area. A combination of these metrics will balance the geographical characteristics in such a way that only those countries will be on the top of the final list which are superior in all areas.

4.2. Technology-based Metrics

Roberts et al. (2011) argue that complexity of a national network infrastructure is defined by two factors – the number of ASs per IP address and the number of IP addresses located away from the core of the network. While the researchers model the first factor as a simple ratio of the number of ASs over the number of IPs in a single country, the second one requires identification of all links between ASs and IPs. Consequently, we suggest extending the geography-based metrics with the following IP-related metrics:

1. The absolute number of IP addresses.
2. The ratio of the number of ASs per IP address.
3. The number of IP addresses per capita.

The first two metrics account for size and complexity of a national network infrastructure. The last one is necessary to better address countries with a low number of IPs as the absolute metric favors countries with high population.

Two online services provide regularly updated statistics about IP addresses on a country level – MaxMind (2013c) and BGPExpert (2013). Both datasets were accessed on July 26th, 2013. This paper uses the statistics collected by MaxMind because of the higher precision of data provided by the service. BGPExpert counts IP addresses per country in millions, significantly decreasing the precision of measurements for small countries.

However, to validate the dataset provided by MaxMind, it was cross-checked with the statistics from BGPExpert for regions with a total number of IP addresses exceeding one million (80 regions in total).

Although an average deviation was only 3.7 %, for some countries results differed significantly, for example, Moldova – 23.5 %, France – 20.5 %, Puerto Rico – 16.4 % and Czech Republic – 11.2 %. The average deviation may be explained by a disparity of the dates when the datasets were last updated – MaxMind's on 07/02/2013 and BGPExpert's on 07/24/2013. The reason for high deviations in some specific cases can also originate from a difference in algorithms of a geolocation process. Notably, about 20 million IP addresses belong to the European Union (EU) in the dataset collected by MaxMind, whereas BGPExpert is more successful in revealing particular countries – only 6 million referred to the general EU region and not to some specific member of the union. In the further analysis the records for the EU and other macro-regions (Asia, Africa, etc.) as well as other unidentified geographical regions (records for anonymous proxy and satellite providers) were excluded from the analysis.

The sanitized MaxMind's dataset contains 246 rows, while the dataset for ASs only 231. The redundant records represent small administrative territories which were not considered in the ranking due to the absence of any ASs within their borders.

4.3. Control-based Metrics

The geographic resilience score should also be extended by including metrics which account for distribution of control over ASs in a country. This control can be considered as control by organizations administrating ASs, Internet service providers (ISPs) and as a political control of a government over ISPs. In the beginning, we chose to use the absolute number of ISPs and the relative number of ASs per ISP in a country as a proxy for the level of control of organizations administrating ASs. The search for a publicly available list of ISPs across countries yielded two sources of information. The first was the World Factbook issued on a yearly basis by the Central Intelligence Agency (CIA) of the United States of America (2013). The second was the online service Whoisthisip.com (2013).

The statistics for ISPs in the World Factbook was available only for the years 2002 and 2004. Further cross-check of the total number of ISPs per country provided by the CIA and the Whoisthisip.com service generated dramatic deviations. Even though the most obvious explanation for this is that the dataset published in the World Factbook is outdated, figures from the online geolocation service were abnormally high. To perform an additional validity test, two countries were selected for a scrutiny. The choice was made in favor of Australia and New Zealand because governmental statistical services of these countries collect and officially publish information about the national network infrastructures including the number of ISPs (Australian Bureau of Statistics, 2012; Statistics New Zealand Tatauranga Aotearoa, 2012). Results of the analysis are presented in Table 1. The disparity in the figures was dramatic, making an extension of the ranking with valid metrics impossible. Therefore, the suggested ISP-based characteristics were not considered in the further analysis.

Country	**The World Factbook**	**Whoisthisip.com (accessed July 26th, 2013)**	**Governmental statistical agencies[5]**
Australia	571	9968	76
New Zealand	36	988	106

Table 1. The total number of ISPs in Australia and New Zealand.

In order to model the political control of a government over ISPs, a metric similar to the one developed by Reynolds and Tamaddon (2011) was included into the ranking. Nonetheless, while their developed metric comprised two factors accounting for censorship and Internet filtering, we propose to use the World Press Freedom Index, published by Reporters Without Borders (2013), as a proxy for a degree of overall freedom that journalists, news organizations, and netizens (active Internet users) experience in a

[5] Both statistical governmental agencies consider only ISPs which provide access to the Internet to more than 1000 customers.

country[6] . The index is based on results of a survey conducted in a form of a questionnaire that was distributed among members of the association, journalists, researchers, jurists, human rights activists and non-governmental partner organizations all around the world. The respondents evaluated a state of the following entities in countries: pluralism, media independence, environment and self-censorship, legislative framework, transparency and infrastructure. A cumulative index assesses media freedom in a single country. In general, the lower the index the higher the freedom of press in a particular country. The last available version of this report was published on 01/30/2013 and contained a ranking of 173 countries. This means that there are 58 missing records in comparison to the dataset for ASs. In the absence of a reasonable approach to fill the gaps and in order not to lose a larger portion of information, a purely technical ranking was calculated apart from the final combined ranking. This ranking considered only the network infrastructure metrics and consisted of 231 countries.

The last metric, suggested for inclusion in the geographical AS resilience ranking, is an average attractiveness of ASs in a country to hackers. The idea behind this metric is that certain parts of a national network infrastructure offer attackers a more significant potential payoff than others. Hence, a risk of becoming a target for attacks related to these segments is higher. Let us consider an example of two countries – one with a prevailing financial sector in an economy and one with a more diversified economy. If the total number of ASs in these countries is similar, then in the first country a large portion of backbone networks will be managed by financial organizations which face frequent hacker attacks (IBM, 2013; Symantec, 2013). As a result, the first country is exposed to a higher risk of an Internet connectivity outage because of simultaneous failures of a bigger number of ASs caused by massive cyber-attacks.

Furnell (2002) defines four components which determine a motivation of hackers: financial payoff, curiosity, notoriety and revenge. An ideal approach for an estimation of the probability of an AS becoming a target of a cyber-attack would include an evaluation of these four components for all ASs. However, this can hardly be performed due to the number of ASs and lack of publicly available information on the data that is hosted on them. This paper proposes a simplified approach which is based on statistics of cyber-attacks collected by major information technology security vendors. The approach includes three distinctive steps:

1. Collection of data on industries which suffered most of the cyber-attacks conducted over the last year.
2. Classification of all ASs by industry.
3. Assignment of a score indicating the probability of becoming a target to each AS depending on the industry it belongs to.

The first step consisted of an analysis of reports published by major IT security vendors. Eight latest of the available reports were downloaded on 19/05/2013 and issued by IBM (2013), Symantec (2013), HP (2012), Cisco (2013), CheckPoint (2013), TrustWave (2013), Trend Micro (2012) and Radware (2012). Only two of them contained information about the distribution of cyber-attacks across industries, and namely the reports provided by IBM and Symantec (see Table 2). Both firms concluded that companies related to the manufacturing, financial and non-traditional (IT & Telecom) services accounted for about 2/3 of all cyber-attacks in 2012. The categories used by Symantec in their report correspond to the Standard Industrial Classification (SIC, 2011), while IBM did not reveal which particular classification they used. Due to this the further analysis was based on the Symantec distribution of cyber-attacks.

[6] The Reporters Without Borders (2013) organization defines the World Press Freedom Index as a measure of "the degree of freedom that journalists, news organizations and netizens enjoy in each country, and the efforts made by the authorities to respect and ensure respect for this freedom".

To complete the second step we used a keyword-based approach to classify ASs by industries for categories concordant to the Standard Industrial Classification System (SIC) (2011). The effective rate of classified ASs reached 13,961/54,773 = 25%. The best result was achieved for African countries – [11] more than 40% of ASs were classified by industry, while for the countries located in Europe and South America the classification rate was only 16%. After the reclassification each AS in the dataset got a specific score based on an industry it relates to. ASs classified into the top three industries and having faced 60% of attacks received the score 3. ASs categorized into governmental, energy/utility and professional service groups got 2. ASs of the rest of industries were assigned 1. All ASs which were left unclassified got the average score of 1.41. Lastly, the resulting dataset was used to calculate an average score for a single country.

Industry	**Share, %**
Manufacturing	26.5
Finance and insurance	20.9
Information and communication	18.7
Health and social services	7.3
Retail and wholesale	6.6

Industry	**Share, %**
Manufacturing	24
Finance, Insurance, Real estate	19
Services – Non-traditional	17
Government	12
Energy/Utilities	10
Services - Professional	8
Aerospace	2
Retail	2
Wholesale	2
Transportation, Communications, Electric, Gas	1

Table 2. Distribution of cyber-attacks across industries (left: IBM, 2013; right: Symantec, 2013).

5. Results and Conclusion

5.1. Results

As already mentioned above, for the analysis a rank was assigned to each country for each of our proposed nine metrics, which positioned the country according to the value of the regarded metric. Finally, all assigned ranks for each country were summed up and based on the sum of ranks a new rank was calculated which constitutes the final rank. Excerpts from the rankings of the ten countries with the most resilient as well as the ten countries with the least resilient networks are displayed in Table 3, left. The ranking which does not consider the results drawn by the Press Freedom Index due to missing information for some countries is shown in Table 3, right.

In summary, mostly European countries are on the top of the ranking. Of all considered European countries, 56 percent can be found at the top quarter of the combined ranking list. No matter what kind of metric is chosen, European countries always dominate the top of the list. North America seems to be somewhere in between. There are resilient countries such as the USA and Canada but also several countries with a low ranking result such as Cuba and Haiti. Considering the size of these well-connected countries, it becomes evident that a large portion of the area of North America is reliably covered. This is caused by the dominant position of the USA and Canada. The same is true for [12] Oceania since Australia and New Zealand – the two best-connected countries in this continental area – account for around 93 percent of the entire region.

In case of Asia, the corresponding countries are almost evenly distributed over the whole combined ranking list (except for the dominant position of the second quarter of the list). It could be the case that there is a trend shifting away from the bottom towards the top, which needs to be checked with historical or future data. Assuming that this is the case, it can be seen as a sign that Asia is already in the advanced state of transition to a more robust Internet connection status. Similarly, South America also seems to be in an early stage of transition. While the top quarter of the combined ranking list contains only three countries from this area, the bottom quarter still accounts for one third of all the South American countries. The shift towards more robustness is observable. African countries fall behind in all areas. More than 65 percent of all African countries can be found at the bottom quarter of the combined ranking list. An exception to this is Gibraltar which has a lot of ASs in relation to its size. This supports the hypothesis that this country is an important connection point for Internet communications. All in all the status quo of each country seems to be closely related to its economic development. Therefore, Africa seems to be still just at the beginning of a period of change.

Position	Nation
1	Latvia
2	Switzerland
3	Romania
4	Poland
5	Austria
6	New Zealand
7	Ukraine
8	Slovenia
9	USA
10	Sweden
…	…
221	Reunion
222	Turkmenistan
223	Cape Verde
224	Chad
225	Ethiopia
226	South Sudan
227	Guinea-Bissau
228	Eritrea
229	Yemen
230	Senegal
231	Korea, D.P.R of (Nord)

Position	Nation
1	Switzerland
2	Latvia
3	Austria
4	Poland
5	Romania
6	New Zealand
7	Sweden
8	Czech Republic
9	Slovenia
10	USA
…	…
169	Congo, The Dem. Rep. Of (Zaire)
170	Burundi
171	Somalia
172	Chad
173	Turkmenistan
174	Guinea-Bissau
175	Ethiopia
176	Senegal
177	Eritrea
178	Yemen
179	Korea, D.P.R of (Nord)

Table 3. Combined ranking for geographical resilience per country with (right) and without (left) considering the press freedom index.

The inclusion of the degree of political freedom into the ranking led to minor changes in positions of countries on the top and in the bottom of the list. Countries best connected to the Internet mostly represent Central, Northern, and Western Europe, while the most vulnerable networks belong to the African and Asian countries.

The aggregated results per continent show that the largest network infrastructure is located in North America and mostly formed by the U.S. and Canada (see Table 4). On a country level Europe represents the least vulnerable region to Internet access disruption. 56 percent of all European countries are in the first quarter of the list. The second largest network infrastructure belongs to Asia. This fact can be explained by the number of users and very rapid spread of Internet technologies in the region. However, the level of freedom is the lowest compared to other continents.

The network infrastructure of Oceania is very similar to that of North America and is mostly constituted by two countries Australia and New Zealand. Although South America, similarly to Asia, looks like a rapidly developing region, the relatively small number of networks and certain problems with political freedom in the region push countries on the continent to the bottom of the list. More than 55 percent of South American countries are in the third quarter. African countries mostly received positions in the bottom of the list. They took 67 percent of positions in the bottom quarter of the ranking.

Continent	# of countries	ASN total	ASN per country	# IP total	# IP per country	Average risk score	Average Freedom Index
Africa	55	968	18	54,129,047	984,165	1.82	35.108
Asia	52	9,835	189	900,677,859	17,320,728	1.82	45.091
Europe	50	17,662	353	677,105,716	13,542,114	1.89	18.921
North America	34	22,338	657	1,697,272,703	49,919,785	1.77	25.267
Oceania	22	1,819	83	55,833,685	2,537,895	1.60	22.649
South America	18	1,352	75	119,957,389	6,664,299	1.90	28.688

Table 4. Aggregated results for geographical resilience analysis.

Further analysis revealed a high correlation of the metrics representing technical characteristics of network infrastructures on a country level. The results of the analysis are summarized in Table 5. For the metrics based on the total number of ASs per country, total number of IPs per country and ratio of the total number of ASs per country over a population density, the correlation was in the range from 0.686 to 0.869. For metrics based on the number of AS per capita and the number of IPs per capita, the correlation reached the value of 0.748. This can be a basis for a simplification of the ranking by exclusion of some metrics to optimize calculations in future work.

	Rank Freedom	Rank Risk Score	Rank Risk Score	Rank AS/IP	Rank #IPs	Rank AS Dens.	Rank AS Pop.	Rank km^2	Rank ASN #
Rank Freedom	1	0.0471	0.523	-0.052	0.142	0.174	0.567	0.457	0.191
Rank Risk Score		1	-0.092	0.151	-0.232	-0.231	0.004	0.043	-0.228
Rank Risk Score			1	-0.436	0.488	0.173	0.748	0.638	0.353
Rank AS/IP				1	-0.748	-0.322	0.121	0.110	-0.400
Rank #IPs					1	0.686	0.014	-0.002	0.869
Rank AS Dens.						1	-0.016	-0.331	0.778
Rank AS Pop.							1	0.821	0.129
Rank km^2								1	0.088
Rank ASN #									1

Table 5. Correlations between the metrics.

5.2. Limitations and Future Work

This study presented a new approach to the resilience assessment of the national network infrastructures. A limitation is that the metrics included into the final ranking so far do not incorporate the existing links between ASs and therefore do not reflect the real complexity of the network structure. Nevertheless, even in the presence of a yet hypothetically appropriate toolset for the revelation of structure, a major problem still involves the identification of all links between ASs. For instance, Dimitropoulos et al. (2007) argue that they were able to discover only 38.7 % of existing peering connections. Furthermore as part for future work it would be additionally possible to include the physical layer in terms of cable maps next to this AS-layer. Another limitation is the lack of reliable statistics on a country level for ISPs. To the best of our knowledge, currently there are no valid services or data sources that can be used for the analysis of a control distribution over national networks.

Moreover, the inclusion of the World Press Freedom Index into the ranking creates a challenge. The reason for this is that not all administrative entities physically hosting ASs are independent countries.

As a solution for this limitation, these territories can be mapped to parent countries and as a result receive the same ranking for this metric. Finally, the applied approach for the evaluation of attractiveness of ASs for hackers was based on the distribution of cyber-attacks in the previous year. Thus, the metric based on this information could be either revised on a yearly basis or calculated using some average figures for a chosen period of time.

5.3. Discussion and Implications

For the global information and network society it becomes increasingly critical to assess how vulnerable specific geographical regions are to Internet access disruptions. Closely related are the options and limits of censorship-based impediments implemented by local governments. For this purpose this paper presented a metric assessing the geographical Internet resilience based on a country level. This metric encompasses several geographic-, technology- as well as control-based indices, combining them into a single, rank-based score that estimates the Internet resilience of a particular country compared to others.

Such a comparison can serve as a systematic assessment for regions of the Internet that are at risk and could support international politics, Internet businesses or freedom activists with indicators for improving Internet resilience and censorship resistance by focusing their activities on critical regions. Moreover, the ongoing debate on net neutrality (Deeb, 2009) would benefit from investigating not only the benefits and drawbacks of providing quality of services to end users but also to investigate geographic factors and political conditions that influence the reliability and speed of accessing a particular content, including risks of not being able to access some content at all.

This ranking will also help to understand and possibly counter geographic options for censorship as well as spying techniques and to assess risks for deploying new services in the network economy. The development and deployment of practical privacy-enhancing technologies can also benefit from insights from our index. Mission critical services should be located in regions with robust and censorship-free network access.

5.4. Conclusion

This paper introduces a proposal for a geographic AS resilience ranking that gives information on the resilience of different countries to Internet disruptions. Our approach broadened the range of factors by including aspects such as the World Press Freedom Index. The final result showed a dominance of European countries in terms of the combination of all metrics, while mostly African countries lack behind in all areas. This confirms to some extent the hypothesis that a national network infrastructure is an outcome of several social, political and technical decisions. All the countries in the bottom of the list refer to regions with poorly developed economics and low level of political freedom. An open question remains what factors play the key role in the development of the network structure on a country level, which can also become a topic for future research.

References

Apnic Aut-Num (2013). Object Template. Available from: http://www.apnic.net/apnic-info/whois_search/using-whois/guide/aut-num/. Last accessed 04/22/2014.

Australian Bureau of Statistics (2012). Internet Activity, Australia. December 2012. Available from: http://www.abs.gov.au/AUSSTATS/abs@.nsf/Latestproducts/FA32C4F74EB0F063CA257B4700137847/. Last accessed 04/22/2014.

BGPExpert (2013). IP Addresses by Country. Available from: http://www.bgpexpert.com/addressespercountry.php. Last accessed 04/22/2014.

Brown, M. (2008a). Ike Brings Biggest Multi-State Internet Outage since 2003. September. Available from: http://www.renesys.com/2008/09/ike-brings-biggest-multistate/. Last accessed 05/06/2014.

Brown, M. (2008e). Pakistan hijacks YouTube. February. Available from: http://www.renesys.com/2008/02/pakistan-hijacks-youtube-1/. Last accessed 05/06/2014.

CAIDA AS Names (2012). CAIDA AS Rank Projekt. Available from: http://as-rank.caida.org/?%20mode0=as-dump-info,%202012. Last accessed 04/22/2014.

CheckPoint (2013). Check Point Security Report. Available from: http://www.checkpoint.com/campaigns/security-report. Last accessed 04/22/2014.

CIA. (2013). The World Factbook. Available from: https://www.cia.gov/library/publications/the-world-factbook/. Last accessed 04/22/2014.

Cisco (2013). Annual Security Report. Available from: http://www.cisco.com/en/US/prod/vpndevc/annual_security_report.html. Last accessed 04/22/2014.

Cowie, J. (2011a). Japan Quake. March. Available from: http://www.renesys.com/2011/03/japan-quake/. Last accessed 05/06/2014.

Cowie, J. (2011b). What Libya Learned from Egypt. March. Available from: http://www.renesys.com/2011/03/what-libya-learned-from-egypt/. Last accessed 05/06/2014.

Cowie, J. (2011c). Egypt Leaves the Internet. Available from: http://www.renesys.com/blog/2011/01/egypt-leaves-the-internet.shtml. Last accessed 05/06/2014.

Cowie, J. (2012c). Syrian Internet is Off the Air. Available from: http://www.renesys.com/blog/2012/11/syria-off-the-air.shtml. Last accessed 05/06/2014.

Cowie, J., Ogielski, A. and Premore, B. J. (2002). Internet Worms and Global Routing Instabilities. Available from: http://www.renesys.com/wp-content/uploads/2013/05/renesys-spie2002.pdf. Last accessed 05/06/2014.

Cowie, J., Popescu, A. and Underwood, T. (2005). Impact of Hurricane Katrina on Internet Infrastructure. http://www.renesys.com/wp-content/uploads/2013/05/Renesys-Katrina-Report-9sep2005.pdf. Last accessed 05/06/2014.

Deeb, K., O'Brien Sr., S.P., Weiner, M.E. (2009). A Survey on Network Neutrality: A new form of discrimination based on network profiling. International Journal of Networking and Virtual Organisations, 6(4):426-436.

Dimitropoulos, X., Krioukov, D., Fomenkov, M., Huffaker, B., Hyun, Y. and Riley, G. (2007). AS Relationships: Inference and Validation. ACM SIGCOMM Computer Communication Review, 37(1): 29-40.

EdgeScape (2013). http://www.akamai.com. Last accessed 04/22/2014.

Furnell, S. (2002). Cybercrime: Vandalizing the information society. Addison-Wesley, Boston.

Hawkinson, J. and Bates, T. (1996). Guidelines for Creation, Selection, and Registration of an Autonomous System (AS). Technical standards document RFC1930, March. Available from: http://tools.ietf.org/html/rfc1930.

HP (2012). Cyber Risk Report. Available from: http://www.hpenterprisesecurity.com/collateral/whitepaper/HP2012CyberRiskReport_0213.pdf. Last accessed 04/22/2014.

IBM (2013). IBM Security Services Cyber Security Intelligence Index. Available from: http://www-935.ibm.com/services/us/en/security/infographic/cybersecurityindex.html. Last accessed 04/22/2014.

ITU (2012). Measuring the Information Society. International Telecommunication Union Report. Available from: http://www.itu.int/en/ITU-D/Statistics/Pages/publications/mis2012.aspx. Last accessed 04/20/2014.

IxMapper. http://www.ixiacom.com/products, 2013. Last accessed 04/22/2014.

Lakhina, A., Byers, J.W., Crovella, M. and Matta, I. (2003). On the Geographic Location of Internet Resources. IEEE Journal on Selected Areas in Communications, 21(6): 934-948.

Madory, D. (2012a). Lebanon Loses Lone Link. July. Available from: http://www.renesys.com/2012/07/large-outage-in-lebanon/. Last accessed 05/06/2014.

Madory, D. (2012b). Blast in Turkey Impacts Iran, Iraq. October. Available from: http://www.renesys.com/2012/10/blast-in-turkey-impacts-iran-i/. Last accessed 05/06/2014.

Madory, D. (2013). Faraway Fallout from Black Sea Cut. February. Available from: http://www.renesys.com/2013/02/faraway-fallout-from-black-sea/. Last accessed 05/06/2014.

Maxmind (2013a). GeoLite Free Downloadable Databases. http://dev.maxmind.com/geoip/geolite. Last accessed 04/22/2014.
Maxmind (2013b). GeoIP Ort - Präzision für ausgewählte Länder. http://www.maxmind.com/de/city_accuracy. Last accessed 04/22/2014.
MaxMind (2013c). Allocation of IP Addresses by Country. Available from: http://www.maxmind.com/en/techinfo. Last accessed 04/22/2014.
National Research Council (2003). The Internet Under Crisis Conditions: Learning from September 11, National Academies Press.
Paul, D., Sarkar, S.K. and Mondal, R. (2013). Optimization of Core Network Router for Telecommunication Exchange. American Journal of Networks and Communications, 2(1): 1-8.
Popescu, A. (2008b). Deja Vu All Over Again: Cables Cut in the Mediterranean. December. Available from: http://www.renesys.com/2008/12/deja-vu-all-over-again-cables/. Last accessed 05/06/2014.
Potaroo (2013). 32-bit Autonomous System Number Report. http://www.potaroo.net/tools/asn32/. Last accessed 04/22/2014.
Radware (2012). Global Application and Network Security Report. Available from: http://www.radware.com/Resources/rclp.aspx?campaign=1630844. Last accessed 04/22/2014.
Renesys (2003). Impact of the 2003 Blackouts on Internet Communications. Available from: http://www.renesys.com/wp-content/uploads/2013/05/Renesys_BlackoutReport.pdf. Last accessed 05/06/2014.
Reynolds, C. and Tamaddon, S. (2011). Network-Political Resiliency. Technical Report. Available from: http://www.chloejreynolds.com/portfolio_materials/NPR-Paper.pdf. Last accessed 04/22/2014.
Roberts, H., Larochelle, D., Faris, R. and Palfrey, J. (2011). Mapping Local Internet Control. Computer Communications Workshop, Hyannis, CA.
SIC (2011). Standard Industrial Classification. Available from: http://www.sec.gov/info/edgar/siccodes.htm. Last accessed 04/22/2014.
Statistics New Zealand Tatauranga Aotearoa (2012). Internet Service Provider Survey. Available from: http://www.stats.govt.nz/browse_for_stats/industry_sectors/information_technology_and_communications/ISPSurvey_HOTPJun12/Definitions.aspx. Last accessed 04/22/2014.
Sterbenz, J. P. G., Hutchison, D., Çetinkaya, E. K., Jabbar, A., Rohrer, J. P., Schöller, M., Smith, P. (2010). Resilience and Survivability in Communication Networks: Strategies, Principles, and Survey of Disciplines. Computer Networks, 54(8), pp. 1245-1265.
Symantec (2013). Internet Security Threat Report 2013. Available from: http://www.symantec.com/security_response/publications/threatreport.jsp. Last accessed 04/22/2014.
Team Cymru (2013). Internet Security Research and Insight. Available from: http://www.cymru.com/. Last accessed 04/22/2014.
TrendMicro (2012). TrendLabs Annual Security Roundup. Available from: http://www.trendmicro.com/cloud-content/us/pdfs/security-intelligence/reports/rpt-evolved-threats-in-a-post-pc-world.pdf. Last accessed 04/22/2014.
TrustWave (2013). Trust Wave Global Security Report. Available from: https://www2.trustwave.com/2013GSR.html?sl=gsr-2012. Last accessed 04/22/2014.
Underwood, T. (2008c). Wrestling With the Zombie: Sprint Depeers Cogent, Internet Partitioned. October 31. Available from: http://www.renesys.com/2008/10/wrestling-with-the-zombie-spri/. Last accessed 05/06/2014.
Wählisch, M., Schmidt, T.C., de Brün, M. and Häberlen, T. (2012). Exposing a Nation-Centric View on the German Internet – A Change in Perspective on AS-Level. Passive and Active Measurement, Springer.
Whoisthisip.com (2013). Available from: http://whoisthisip.com/internt-service-provider-isp-ip-usage.php. Last accessed 04/22/2014.
Wikipedia ISO 3166 Coding List (2013). ISO-3166-1-Kodierliste. http://de.wikipedia.org/wiki/ISO-3166-1-Kodierliste. Last accessed 04/22/2014.
Wilcox, S. (2007). Quaking Tables – The Taiwan Earthquakes and the Internet Routing Table. http://www.renesys.com/wp-content/uploads/2013/05/ripe-quaking-tables.pdf, 2007. Last accessed 05/06/2014.Reporters Without Borders (2013). World Press Freedom Index. Available from: http://en.rsf.org/press-freedom-index-2013,1054.html. Last accessed 04/22/2014.
Zmijewski, E. (2008d). Georgia Clings to the 'Net. August. Available from: http://www.renesys.com/2008/08/georgia-clings-to-the-net/. Last accessed 05/06/2014.

ARTICLE 4:

WHO RUNS THE INTERNET? - CLASSIFYING AUTONOMOUS SYSTEMS INTO INDUSTRIES[7]

Authors: Annika Baumann, Humboldt-Universität zu Berlin, Berlin, Germany

Benjamin Fabian, Humboldt-Universität zu Berlin, Berlin, Germany

Proceedings: 10th International Conference on Web Information Systems and Technologies (WEBIST'14)

Publication year: 2014

Abstract

The Internet consists of a network of Autonomous Systems (ASs). To understand which kind of organizations control those ASs can help to better assess the Internet structure in terms of economic interests and reliability. The current paper proposes a novel classification approach by combining AS-specific data with business data from the United States Securities and Exchange Commission. Furthermore, more detailed industry classes than in previous works are considered, inspired by the North American Industry Classification System (NAICS). Using our methodology on a recent data set, we were able to classify 56.69 % of the considered ASs into industries. This lays a foundation for our future work on investigating the important players of the Internet backbone as well as their economic interests and risks.

***Keywords**: Internet, Autonomous System, Industry, Classification*

[7] This article is provided with kind permission from SciTePress. The original version is available at: https://doi.org/10.5220/0004936803610368.

1. Introduction

The Internet expanded rapidly during the last decade. From 2001 to 2013, the fraction of the world population using the Internet increased from 8.0 % (International Telecommunication Union, 2011) to an estimated 38.8 % (International Telecommunication Union, 2013), with a simultaneous population increase from 6.1 billion (United Nations Population Fund, 2001) to 7.2 billion people (U.S. Department of State, 2013), resulting in approximately 2.8 billion Internet users today versus 0.49 billion in 2001. This rapid growth in users resulted in a heterogeneous and complex system, making analysis and modelling of the Internet structure difficult.

Our paper is part of an on-going research project that is investigating how the Internet of today is structured in terms of economic interests, control and reliability. Who are the important players of the Internet backbone, what are their economic interests and risks with respect to their business models, and what are the implications for reliability, security and privacy as well as political control?

Our first step towards approaching these goals is to classify the important organizations that control Autonomous Systems (ASs) of the Internet according to business categories, which could support future analyses along all of those dimensions. For example, with respect to reliability and security, common methods assess the robustness of the Internet structure based on graphs and modelling the Internet as an abstract complex network consisting of nodes (each representing an AS) that are connected via edges. However, such approaches solely focus on topology-based robustness and so far ignore the highly economically driven character of the Internet, as well as corresponding heterogeneous risks of attack and control.

At an organizational and global routing level of abstraction, the Internet can be considered as composed of ASs. An AS can be defined as “a group of IP networks run by one or more network operators with a single clearly defined routing policy; when exchanging routing information to the outside, each AS is identified by a unique number (Réseaux IP Européens, 2011). The *Internet Corporation for Assigned Names and Numbers* (ICANN) and, via delegation, the *Regional Internet Registries* (RIR) are responsible for registration of these AS numbers (ASNs). The amount of registered ASs increased from roughly 10,000 in the year 2000 to more than 60,000 in 2013 (Potaroo, 2012), which is also another indicator for the substantial increase of Internet complexity.

Classifying the major players of the Internet backbone is an interesting challenge in itself because publicly available business data is sparse. Our approach presented in this article focuses on analyzing the public registration information for AS numbers. Moreover, we present an approach for the classification of ASs into detailed industry classes in order to better understand the organizational and economic patterns of the Internet.

The rest of the paper is structured as follows: Section 2 discusses related work. Section 3 presents the data sources, followed by Section 4 on our methodology. Section 5 presents our results, and Section 6 concludes the paper.

2. Related Work

Some earlier research articles proposed approaches for classifying ASs into various categories. The classification approach used in our paper was initially inspired by the methods employed by Dimitropoulos et al. (2005). Based on an expert system that uses text classification techniques, the authors used organization names to categorize ASs. Each AS was assigned to one or more of the basic classes Internet service providers (ISP), Internet exchange points (IXP), network information centers (NIC), companies providing no Internet service as well as education- and research-, military-, government- and health-related networks. The authors were able to classify 20,598 out of 32,689 ASs in 2005, which corresponds to 63.01%.

Another work (Dimitropoulos et al., 2006) used even more coarse-grained classification categories, namely only large and small ISPs, customer ASs, universities, IXPs and NICs. The method applied was based on the AdaBoost algorithm (Freund and Schapire, 1997) using several attributes (e.g., organization description; number of inferred providers, customers and peers; number of advertised IP prefixes) to classify the relevant ASs into their respective classes. The authors were able to classify 95.3% of 19,537 ASs with an accuracy of 78.1%.

The main focus of the work by Chang et al. (2005) was to estimate traffic volume between individual ASs. For this, the authors classified ASs regarding their initial utility, which resulted into the three classes web hosting, residential access and business access. The methodology used by the authors is different from other work conducted in this area. Instead of investigating an individual AS and assigning it to a class, they created a class and tried to find relevant ASs on the Internet. The authors were able to identify 56% of all BGP-advertised ASs with their approach.

The primary focus of the paper by Dhamdhere and Dovrolis (2011) was to analyze the evolution of the AS ecosystem over the last 12 years. ASs were classified into the classes enterprise customers, small and large transit providers, access/hosting providers and content providers. A decision tree approach was applied for classification. In order to build the training set, for each class 50 ASs were classified manually. Afterwards, the classification was conducted for 42,000 ASs by using the number of customers and the number of peers as independent variables. Classification accuracy for the classes ranged between 76% and 82%.

All of those articles have in common that the proposed classes are not comprehensive and do not resemble real industries. Thus they contribute not much to a better understanding of the industry structure behind the ASs comprising the Internet. Our work addresses this research gap by proposing a classification approach that adopts fine-grained industry classes.

3. Data Sources

3.1. CAIDA

The Cooperative Association for Internet Data Analysis (CAIDA) "is a collaborative undertaking among organizations in the commercial, government, and research sectors aimed at promoting greater cooperation in the engineering and maintenance of a robust, scalable global Internet infrastructure." (CAIDA, 2011). One project offered by CAIDA is the *AS Rank* project (CAIDA, 2012). It is based on Border Gateway Protocol (BGP) routing data collected by RouteViews (2013) and the RIPE NCC (2013). The list of ASs that is used in our paper contains the information of 59,576 ASs. An excerpt of the dataset can be seen in Figure 1. For the purpose of classifying ASs into industry classes mainly the *org name* attribute was considered as highly relevant.

```
# format: AS number|source|AS name|country|org name|org_id|date
1|ARIN|LVLT-1|US|Level 3 Communications|LVLT-ARIN|20120224
```

Figure 1. Excerpt of CAIDA AS Rank

3.2. SEC

The U.S. Securities and Exchange Commission (SEC) is a government agency in the USA (United States Securities and Exchange Commission, 2013). Its primary purpose is to regulate securities and enforce federal securities laws. Every company publicly traded in the United States has to file certain documents with the SEC. The Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system makes those filings available to the public. This can be used to gather the Standard Industrial Classification (SIC) code for the company (Figure 2). An SIC code can be directly mapped to an NAICS code using a mapping table (CareerOneStop U.S., 2013). Thus it is possible to uniquely identify the industry of an AS's organization by use of the EDGAR system. A limitation is that only organizations that are listed on a stock exchange in the USA can be found in the system.

INTERNATIONAL BUSINESS MACHINES CORP
SIC: 3570 - COMPUTER & OFFICE EQUIPMENT
State location: NY | State of Inc.: **NY** | Fiscal Year End: 1231

Figure 2. Excerpt of SEC EDGAR result.

3.3. RIR AS Information

As an additional information source, data from the RIRs was retrieved. The website *cidr-report.org* contains AS information from all RIRs. It allows searching for individual ASs and returns the information that comes from the WHOIS services of the individual RIRs. In order to simplify the data retrieval process, this website was also used to retrieve AS-specific WHOIS information instead of using the WHOIS services of different RIRs. A sample of such information can be seen in Figure 3.

```
aut-num:        AS6619
as-name:        SAMSUNGSDS-AS-KR
descr:          SamsungSDS Inc.
descr:          Seoul Yeoksam-dong Gangnam-gu   707-19
descr:          135-080
country:        KR
```

Figure 3. Sample of RIR AS information.

4. Methodology

Figure 4 gives an overview of the process of classifying the ASs presented in this paper. As a first step, the relevant industry classes for the classification approach needed to be defined. Their definition draws from the North American Industry Classification System (NAICS, 2013). Due to the intrinsic online setting of our investigation, special adjustment was necessary, meaning that several of these classes were either merged, dropped or changed. In the case of ASs, some industries are missing at all while some of them are overrepresented. Therefore, the NAICS was only used as a basis for the classification approach in our particular setting. An overview of the classes can be found in the Appendix.

Step 1: Preprocessing. The initial AS list included data from the year 2012 and was taken from the CAIDA AS Rank project (CAIDA, 2012); it contained 59,576 ASs. In order to only include reasonable and recent data, the list was preprocessed. At first, the information gathered from the RIRs was used to filter for inactive ASs. This reduced the list by 17,830 ASs, leaving 41,746 ASs to classify. Furthermore, all ASs that did not have an according organization name, i.e., all entries either containing no specification of the underlying organization or being a no registry entry, were removed from the list. Eliminating 1,362 ASs, this step left 40,384 ASs in the list.

Step 2: Keyword Classification. In the next step, a keyword list was created by analyzing word and phrase frequencies with the help of an occurrence counting of words, bi-grams and tri-grams. All words and phrases that appeared quite frequently were analyzed in more detail. It was assumed that tri-grams needed to occur at least five times, bi-grams ten times and simple words twenty times to be selected for deeper analysis. The rationale behind this procedure was to include only those words and phrases that are most frequent and therefore important. This makes it possible to classify several ASs at the same time based on a single phrase or keyword. Keywords were mainly defined in such a way that the organization name or a part of it had to comply with the complete keyword. This means that for example in case of the keyword "ship" only the word itself would fit and not "membership" or "ownership". This was done to ensure the reliability of keywords by avoiding undesired mismatches. The selection of keywords itself was randomly cross-checked based on real data to further ensure their reliability and unambiguity. Only those words or phrases were chosen whose unambiguity in relation to industry classification was satisfactory. For example the keyword "Internet service provider" is highly reliable if it comes to sorting into the category ISPs & Networks, while "service provider" might lead to wrong results for the same category. Organizations having a (part of their) name such as "content service provider" would also fit into such a category.

In order to minimize wrong categorizations, an iterative learning process was applied. The procedure was as follows: based on the first selection of keywords, the AS numbers were categorized into the industry classes created so far. Each category was then checked for wrong categorizations. For this purpose, the list of categorized ASs and their underlying organization was reviewed manually. If the categorization of an ASN was wrong, the reason was identified and eliminated with the help of refined or discarded keywords. This procedure helped to ensure that only those keywords remained that are at the same time reliable and general. In order to check for further yet not identified keywords, a list was generated that contained all non-categorized ASNs. This list was then manually checked for further keywords at each iteration. This was particularly important in case of misspelling and language-specific variations. For example, the keyword "university" was represented by many language specific variations such as "universitas", "universidad" or "univ". An example for misspelling is "network *infomation* center" which occurred at least seven times in the list. Such variations were additionally included in the keyword list for each category.

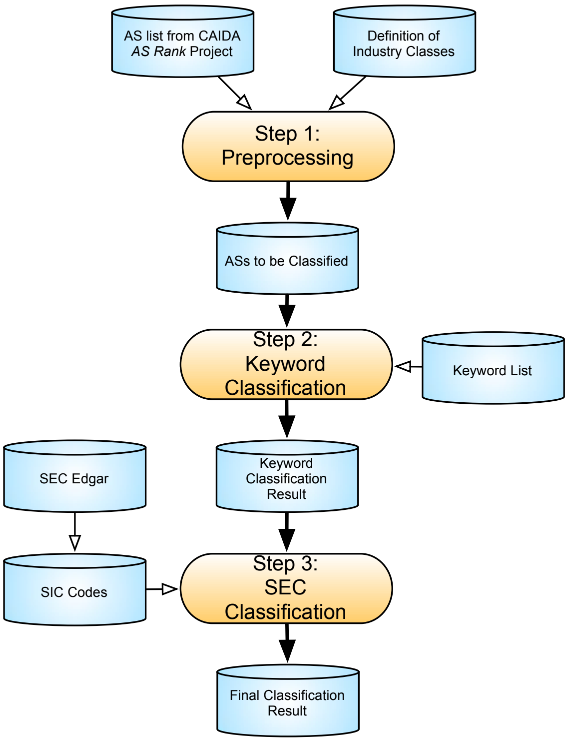

Figure 4. *Process of AS industry classification.*

Based on this extended and refined keyword list, the procedure started from the beginning and was repeated again. The complete list of the industry classes created and their respective definition are shown in Table 1. The keywords used for each industry class are given in the Appendix.

Cluster	Definition	Result	Ratio
Address	ASs where no underlying organization is specified but an address, where the AS itself, the underlying organization or its managerial unit is located.	326	0.81%
Company	Collecting bucket for those ASs which are hard to categorize based on their organization name but at least can be identified as a company.	9,104	22.54%
Construction & Manufacturing	Mostly building firms and manufacturers are part of this class.	97	0.24%
Consulting & Management	ASs related to advising and leading of company.	139	0.34%
Education & Research	ASs related to learning and gaining of new insights such as schools, universities, research facilities and networks as well as laboratories.	2,150	5.32%
Entertainment & Information	ASs which are for example related to television, gaming, radio or publishing.	761	1.88%
Finance & Insurance	This class consists mainly of banks and insurance firms.	1,664	4.12%
Government & Military	ASs with an authority and military character as well as areal territories such as cities and states are relevant for this class.	1,039	2.57%
Healthcare	Next to hospital (district) related entities, this class contain pharmaceutical firms.	695	1.72%
IT & Internet Service	ASs that are affiliated with online as well as offline IT services and computer products. In general, this includes those firms which provide a service or product that is based on the Internet or IT, but which do not offer Internet access.	1,371	3.39%
ISPs & Networks	Collects ASs of those organizations which offer Internet access or provide the necessary infrastructure.	3,775	9.35%
IXPs	This class collects all ASs which function as exchange point in the Internet.	134	0.33%
NIC	Contains those ASs which are "responsible for managing and allocating Internet resources" ().	390	0.97%
Space	Contains ASs of the area of astronautics.	30	0.07%
Telephone & Communication	This class contains (mobile) telephone providers and sellers as well as general communication-based organizations.	2,118	5.24%
Travel	Contains all ASs that are related to mobility and travel, such as airports, train stations, hotels and travel agencies.	103	0.26%
Trade & Transport	Collects ASs of the area of wholesale and logistics including apparel and food.	232	0.57%
Utilities	Organizations which provide electric power, water as well as other basic materials; also services such as waste disposal, coal and mining belong to this class.	256	0.63%

Table 1. Industry classes with their definitions and percentages based on the keyword classification.

Step 3. SEC Classification. A Java program was written to download information from the SEC EDGAR system. The organization name was used to search for the company. For 40,384 search requests, 2,732 entries could be found in the EDGAR system. However, sometimes the same company has several names, which resulted in more than one outcome for the organization name. An example of such an ambiguity can be found in Figure 5. Because there was no reliable way to uniquely identify the correct entry in such a case automatically, all entries with multiple search results were eliminated which led to 1,706 remaining search results. Furthermore some companies had no SIC code and were eliminated as well. This resulted in 469 ASs that could additionally be classified into industry groups.

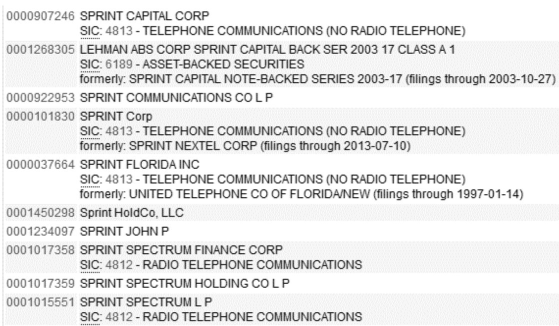

Figure 5. Ambiguous EDGAR search result for "Sprint".

5. Results

5.1. Keyword Classification

Applying the method described above and using the keywords shown in the Appendix to classify the 40,384 ASs, resulted in 22,786 or 56.42 % of classified ASs. The industry class distribution based on keyword classification only can be seen in Figure 6. According to this data, most frequently the organizations that own ASs belong to the industry classes *Education & Research*, *Finance & Insurance*, *ISPs & Networks*, and *Telephone & Communications*. This class distribution seems to be intuitive: ISPs, telephone and IT companies as well as universities have more incentives to register an AS than for example a travel agency because ASs classified into these categories are often related to communications, but often also represent major institutions that have a high tendency to own an AS simply because of their size.

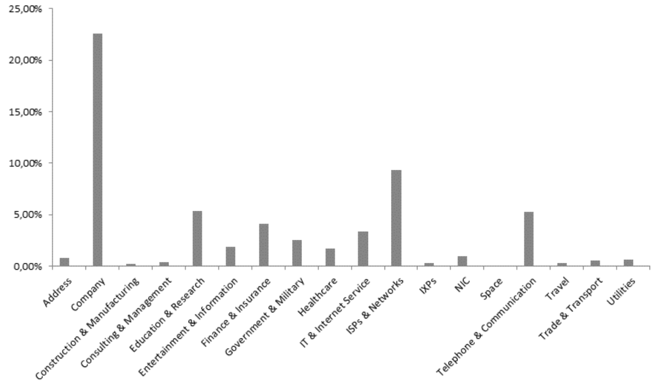

Figure 6. Industry class percentages based on keyword classification.

Of all clusters (apart from the generic *Company* cluster) *ISPs & Networks* is the category that is most frequent and accounts in case of both classification approaches for around ten percent of all the classified entities. This is an expected result since ASs pertain to the communications business, and offering Internet access is a key factor in this particular business area. The categories *Education & Research* and *Telephone & Communication* occupy the second and third positions with 5.32 % and 5.24 % respectively. Therefore, even today, companies in the area of Internet and information technology are still overrepresented because of their particular Internet affinity. A bit more unexpected, however, is that financial institutions seem well represented since the fourth position is taken by the *Finance & Insurance* cluster. All other categories are smaller with percentage values between 2.57 % (*Government & Military*) and 0.07 % (*Space*).

However, a limitation of our results is that the general *Company* cluster still encompasses 22.54 % of the classified ASs. This fact and the remaining number of unclassified ASs indicate that there is still a potential for improvement regarding the classification process. Yet it is questionable whether it is possible to reach much better results with semi-automatic classification approaches because of the presence of non-self-explanatory organization names and acronyms such as NGM or EDP. Not only is it difficult to classify those simply based on keywords, it is also challenging to specify what kind of organization they represent without further manual and individual investigations.

5.2. SEC Classification

The industry-class frequency of organizations based on an alternative classification that is solely based on SEC data is shown in Figure 7. The industry classes *Construction & Manufacturing* as well as *Consulting & Management* have the most organizations with ASs. The classes *Government & Military*, *ISPs & Networks*, *IT & Internet Service*, *IXPs*, *NIC*, as well as *Space* have no ASs at all. However, because not all companies are listed with the SEC and in particular governmental institutions and privately held companies are not registered, the lack of representation of these classes is inherent.

With the help of the SEC data it was possible to classify additional 116 ASs, which could not be classified via keywords only (Figure 8, left). Furthermore, the industry classes of 206 ASs could be specified more precisely which had previously been assigned to the *Company* class (Figure 8, right). Most newly specified classifications were assigned to the *Construction & Manufacturing* as well as the *Consulting & Management* industry classes. Their prevalence reflects the results of the SEC classification.

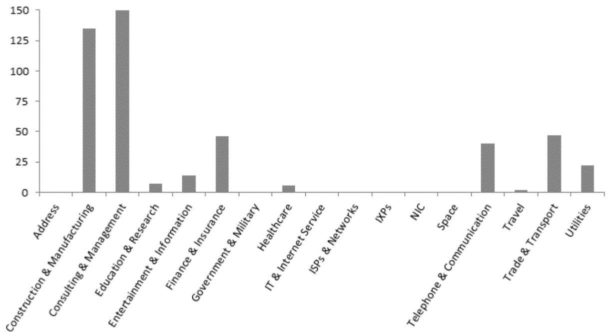

Figure 7. Industry class sizes based on SEC classification.

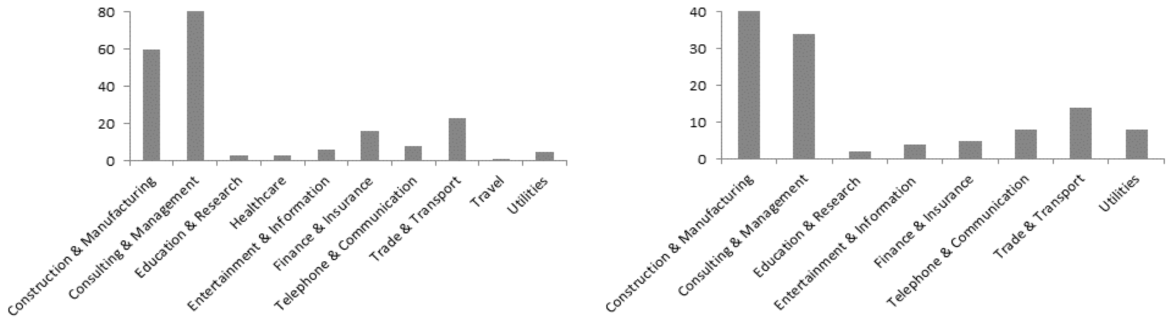

Figure 8. New classifications (left) and more precise SEC classification (right).

By combining both classification approaches we were therefore able to classify 22,892 of the 40,384 AS of the preprocessed list. This accounts for 56.69 % of all considered ASs that could be assigned to an industry class. The final result is shown in Table 2.

Cluster	**Result**	**Ratio**
Address	326	0.81%
Company	8,898	22.03%
Construction & Manufacturing	198	0.49%
Consulting & Management	254	0.63%
Education & Research	2,155	5.34%
Entertainment & Information	771	1.91%
Finance & Insurance	1,685	4.17%
Government & Military	1,039	2.57%
Healthcare	698	1.73%
IT & Internet Service	1,371	3.39%
ISPSs & Networks	3,775	9.35%
IXPs	134	0.33%
NIC	390	0.97$
Space	30	0.07%
Telephone & Communication	2,134	5.28%
Travel	104	0.26%
Trade & Transport	2,546	6.30%
Utilities	269	0.67%

Table 2. Final classification using both keyword and SEC data (Note: It was possible to categorize an AS into more than one industry class).

6. Conclusions

This paper proposed a classification approach for categorizing ASs into detailed industry classes in order to better understand the economic background of the Internet structure. The industry classes are inspired by the NAICS (2013), which had the effect that an unprecedented level of detail regarding the industry classes for classification could be achieved. Data was mainly obtained from the CAIDA AS Rank project as well as from SEC.

The classification of ASs into industry classes based on their underlying organization revealed an on-going strong dominance of telecommunication and IT-related firms in the current Internet as well as of large institutions such as banks and universities.

It was possible to classify 56.69 % of all ASs (after preprocessing). Nevertheless, the amount of unclassified ASs indicates that there is room for improvement regarding the categorization process. A refined and extended keyword selection process could provide better results. Nevertheless, since there is a non-negligible amount of ASs having organizational specifications that are not self-explanatory or acronyms, this would involve a difficult challenge.

Some of our further explorative attempts to find new ways for AS classification with the help of clustering algorithms had limited success so far. However, another possible route could be to apply methods from Natural Language Processing (NLP) to the AS data and also for analyzing search results from the Web for acronyms or other challenging organization names.

Moreover, customers of the various ISPs cannot be captured by the current method. It is often the case that large Internet providers also represent smaller customers who are not registered in the organizational information of the ASs. Here, studying the level of IP addresses could provide further insights but will also involve complex challenges.

Various other classification approaches might be feasible. In future work we will try to find other valuable classification systems aiming to take an even closer look at the composition of the Internet. Furthermore, we will use our classification results to further investigate the important players of the Internet backbone as well as to assess their economic interests and risks, at individual as well as global scales. Moreover, we aim to derive implications for Internet reliability and control assessments as well as for security and privacy analyses.

Acknowledgements

The authors thank Sebastian Dombrowski for his programming work during parts of this research.

Appendix

Cluster	Keywords
Address	avenue, building, flat, floor, strasse, gpo box, handelsweg, mcpo box, no., po box, road, street, suite(s), tower
Company	associates, agency, a\.s\., bv, b\.v\., cjsc, co, co kg, companies, company, coporation, corp, corporation, d.o.o., de c.v., enterprise(s), gmbh, inc, incorporated, l\.l\.c, limited, llc, llp, lp, l\.p\., ltd(a), organization, s.a. de c.v., s.p.a., s.r.l., sp. z o.o., srl, s\.a\., s\.l\., sa, sas, sl, trust, z\.s\.p\.o, zspo
Construction & Manufacturing	architect(s), builders, building company, building society, construcoes, construction, constructora, constructors, electronics, machine, manufacturer(s), manufacturing, producers
Consulting & Management	beratung, business solutions, capgemini, consultance, consultancy, consultant(s), consulting, ernst & young, management company, pricewaterhousecoopers
Education & Research	.*universitaet, academic, academisch, colegio, college(s), desire2learn, ecole, education(al), fachhochschule, forschungsgemeinschaft, forschungsgesellschaft, fraunhofer, institute, instituto, knowledge network, laboratories, laboratory, labs, learning, mitre, physics, polytechnic, recherché, research, school(s), science(s), supercomputer, supercomputing, univ, universidad, universidade, universitaet, universitaria, universitas, universite, universiteit, universitesi, universitet, universiti, universities, university, univerzitet
Entertainment & Information	advertising, bbc, bertelsman, book(s), broadcasting, entertainment, football, fun, game, gaming, library, magazine(s), mcgraw-hill, media, marketing, medien, multimedia, news, newspaper(s), printing, publications, publishing, radio, reuters, television, times, tv, weather, zdf
Finance & Insurance	allianz, american express, asset management, assurance, banca, banco, bank, banka, banque, blue shield, capital, credit, finance, e*trade, financial, goldman, guggenheim, hsbc, insurance, investment, leasing, payment, real estate, reinsurance, rental, societe generale, stock exchange, stonepeak, visa
Government & Military	administration of, agency, air force, army, authority, board of, bureau of, city of(fice), committee, commonwealth of, congress(ional), council, county of(fice), department of, dept, district of(fice), dod, embassy, federated states, gov, government, house of, iles de, military, ministry, nato, navy, northrop grumman, parliament(ary), province of, senate, state of, united nations, united states postal service, US geological survey
Healthcare	bayer, blood, dental, drug(s), drugstore, elektromedizinische, emergency, health(care), hospital(s), johnson & johnson, klinikum, medical, medicine, medizinische, merck, novartis, pfizer, pfizerswitzerland, pharma(cy), pharmaceuticals, pharmafarm, propharma, social security, transplant
IT & Internet Service	akamai, apple inc, computer hardware, computer products, computer science, computer service(s), computer software, computer solutions, computer systems, content provider, content service provider, content solution(s), data center(s), data corporation, data processing, data service(s), data solution(s), data systems, dell, fujitsu, general electric, google, hewlett-packard, host, hosting, ibm, information systems, information technology, internet service(s), internet systems consortium, it services, microsoft, neterra, network service(s), network systems, oracle, othello, samsung, sap, schuberg philis, siemens, sony, sungard availability, thinktech, verisign, web service(s), yahoo
ISPs & Networks	aol, arcor, at&t, backbone, broadband, bt italia, cable network(s), cogent, comcast, connection(s), esnet, exatel, fibernet, freenet, gts, iletisim hizmetleri, internet access, internet provider, internet service provider(s), internet solution(s), isp, lattelekom-apollo, level 3, linxtelecom, netassist, netcologne, network access, network provider, network service(s), network solution(s), networks, ntt america, optical network, prometey, qwest communication(s), reseau national, reseau regional, retn, road runner, rostelecom, singtel optus, smartcity, sprint, surfnet, swisscom, t-2, telecom, telekom, telia latvija, teo, time warner cable, towerstream, transit, true internet, uzbektelecom, verizon, versatel, vimpelcom, west call, wireless
IXPs	exchange point, internet exchange, internet exchange point, ix, ixp(s), link, open exchange, peering exchange
NIC	afnic, american registry, apnic, arin, east-ukrainian, internic, network information center, network information center, network information centre, nic, ripe ncc
Space	aeronautics, aerospace, astronomy, nasa, space administration, space agency, space research, space telescope
Telephone & Communication	alcatel, bell canada, communication(s), e-plus, elisa, ericsson, lambdarail, mobile, motorola, nokia, o2, phone, radiotelephone, rockefeller group, singtel optus, telecommunication(s), telecomunicaciones, telefonica, telekommunikation, telekomunikacije, telekomunikacja, telekomunikasi, telecomunicazioni, telephone(s), telianet, turkcell, vodafone
Travel	air canada, airline(s), airport, bahn(hof), boing, flughafen, klm, lufthansa, hotel(s), railway, reisebuero, resort, travel, vacation
Trade & Transport	amazon, apparel, clothing, coca-cola, fedex, food(s), logistic(s), logisticare, retail(ers), shaya magazacilik, shipping, shoe(s), supply, trade, trading, transport(ation), wal-mart, wholsesale
Utilities	bp, coal, electric power, electricity, energy, farmer, farms, fiber, gas, mine, mining, offshore, petroleum, utilities, utility, waste, water

Table A. Keywords used for each industry class.

References

CAIDA, 2011. About CAIDA. http://www.caida.org/home/about/. (Access Dec, 2013).

CAIDA, 2012. AS Rank Project. http://as-rank.caida.org/?mode0=as-dump-info. (Access Dec, 2013).

CareerOneStop U.S. Department of Labor, Employment and Training Administration, 2013. NAICS-SIC Cross/Reference. http://www.acinet.org/industry/Ind_Sic.aspx?id=&nodeid=1. (Access Dec, 2013).

Chang, H., Jamin, S., Mao, Z., Willinger, W., 2005. An Empirical Approach to Modeling Inter-AS Traffic Matrices. In *Proceedings of the 5th ACM SIGCOMM Conference on Internet Measurement*.

Dhamdhere, A., Dovrolis, C., 2011. Twelve Years in the Evolution of the Internet Ecosystem. *IEEE/ACM Transactions on Networking* 19(5):1420–1433.

Dimitropoulos, X., Krioukov, D., Riley, G., claffy, k, 2005. Classifying the Types of Autonomous Systems in the Internet. *SIGCOMM 2005 Poster*, Philadelphia, Pennsylvania.

Dimitropoulos, X., Krioukov, D., Riley, G., claffy, k. 2006. Revealing the Autonomous System Taxonomy: The Machine Learning Approach. In *Passive and Active Network Measurement Workshop (PAM)*, Adelaide, Australia.

Freund, Y., Schapire, R.E., 1997. A Decision Theoretic Generalization of Online Learning and an Application to Boosting. *Journal of Computer and System Sciences* 55(1):119–139.

International Telecommunication Union, 2011. Internet Users per 100 Inhabitants 2001-2011. http://www.itu.int/ITU-D/ict/statistics/material/excel/2011/Internet_users_01-11.xls. (Access Dec, 2013).

International Telecommunication Union, 2013. Key ICT Indicators for Developed and Developing Countries and the World. http://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2013/ITU_Key_2005-2013_ICT_data.xls. (Access Dec, 2013).

NAICS, 2013. North American Industry Classification System. http://www.census.gov/eos/www/naics/. (Access Dec, 2013).

Potaroo, 2012. 32-bit AS Number Report. http://www.potaroo.net/tools/asn32/. (Access Dec, 2013).

RouteViews, 2013. University of Oregon Route Views Project. http://www.routeviews.org/. (Access Dec, 2013).

Réseaux IP Européens, 2011. AS Number Assignment Policies. http://www.ripe.net/ripe/docs/ripe-525. (Access Dec, 2013).

RIPE NCC, 2013. Réseaux IP Européens Network Coordination Centre. http://www.ripe.net/. (Access Dec, 2013).

U.S. Department of State, 2013. World Population Day 2013. http://www.state.gov/secretary/remarks/2013/07/211828.htm. (Access Dec, 2013).

United Nations Population Fund (UNFPA), 2001. The State of World Population 2001 – Demographic, Social and Economic Indicators. https://www.unfpa.org/swp/2001/english/indicators/indicators2.html. (Access Dec, 2013).

United States Securities and Exchange Commission (SEC), 2013. The Investor's Advocate: How the SEC Protects Investors, Maintains Market Integrity, and Facilitates Capital Formation. http://www.sec.gov/about/whatwedo.shtml. (Access Dec, 2013).

ARTICLE 5:

EXPLORING THE BITCOIN NETWORK[8]

Authors: Annika Baumann, Humboldt-Universität zu Berlin, Berlin, Germany

Benjamin Fabian, Humboldt-Universität zu Berlin, Berlin, Germany

Matthias Lischke, Humboldt-Universität zu Berlin, Berlin, Germany

Proceedings: 10th International Conference on Web Information Systems and Technologies (WEBIST'14)

Publication year: 2014

Abstract

This explorative paper focuses on descriptive statistics and network analysis of the Bitcoin transaction graph based on recent data using graph mining algorithms. The analysis is carried out on different aggregations and subgraphs of the network. One important result concerns the relationship of network usage and exchange rate, where a strong connection could be confirmed. Moreover, there are indicators that the Bitcoin system is a "small world" network and follows a scale-free degree distribution. Furthermore, an example of how important network entities could be de-anonymized is presented. Our study can serve as a starting point in investigating anonymity and economic relationships in Bitcoin on a new structural level.

Keywords: *Electronic Cash, Bitcoin, Network Analysis*

[8] This article is provided with kind permission from SciTePress. The original version is available at: https://doi.org/10.5220/0004937303690374.

1. Introduction

Bitcoin is a decentralized digital currency based on a peer-to-peer network architecture and secured by cryptographic protocols. It was originally proposed by Nakamoto (2009). Anonymity and avoidance of double spending are realized via a block chain, a kind of transaction log that contains all transactions ever carried out in the network. In order to provide some anonymity, personal, identifiable information is omitted from the transaction. Therefore, the source and destination addresses are encoded in the form of public keys. Every public key, which serves as pseudonym, has a corresponding private key which is stored in an “electronic wallet”. Private keys are used to sign or authenticate any transactions. To become part of the peer-to-peer network, one needs to install a client software that runs either on a local device or at cloud providers (Ober, Katzenbeisser, Hamacher, 2013).

Authorization and verification are conducted by a complex proof-of-work procedure. Nakamoto (2009) proposed the use of a timestamp server which takes the hash of a block of items, timestamps it, and widely publishes the hash to the network. The proof-of-work also creates new Bitcoins in the network; this process is called “mining”. Creation of Bitcoins is limited to a fixed amount of 21 million Bitcoins that can be introduced to the system; this limitation aims at avoiding inflation. Therefore, until that point is reached around the year 2140, money supply will increase at a certain rate every year (Drainville, 2012).

Our explorative work applies descriptive statistics and network analyses to the Bitcoin transaction graph. The network data was provided by Brugere (2013) who applied several tools for downloading and constructing the user network of Bitcoin. Several aggregations are used to highlight network characteristics. The research focuses on global time-varying dynamics within the network. As a first step of our methodology, qualitative research was conducted in order to gain an insight into related work and the transaction graph. Next, we explored the provided data and undertook required preprocessing steps for storing it appropriately in a database. Statistics and network analyses were conducted using this database; results were evaluated, interpreted, and compared to recent research on the transaction graph.

2. Related Work

There are three main related research articles on the Bitcoin transaction graph that were published within the last two years. The most recent work carried out by Ober, Katzenbeisser and Hamacher (2013) focuses on time-varying dynamics of the network structure and the degree of anonymity. Using data of the period 03/01/2009 to 06/01/2013, the authors discovered that the entity sizes and the overall pattern of usage became more stationary in the last 12 to 18 months, which reduces the anonymity set. The authors also show that the number of dormant coins is important to quantify anonymity. Inactive entities hold many of these dormant coins and thus further reduce the anonymity set (Ober, Katzenbeisser, Hamacher 2013).

Reid and Harrigan (2013) focus on anonymity in the Bitcoin network, analyzing the topology of the transaction and user network based on data of the time interval from 03/01/2009 to 12/07/2011. The authors adopt a preprocessing step to construct the user network. In order to improve the anonymity analysis, the researchers propose several methods including the integration of external information that is mainly held by businesses and other services which accept Bitcoin as payment. They show that it is possible to associate IP addresses from a public service with the recipient’s public keys and link it to previous transactions.

In the third paper by Ron and Shamir (2013) the main focus lies on non-dynamic statistical properties of the transaction graph. The authors analyzed data of the period from 03/01/2009 to 13/05/2012, using various statistics such as distributions of addresses, incoming BTCs, balances of BTCs, number and size of transactions, and most active entities. They found that the majority of Bitcoins is not in circulation and that most of the transactions amount to a rather modest sum (less than 10 BTC). The researchers also analyzed the largest transactions in the network (greater than 50,000 BTCs) and determined their flows. They showed that most of these transactions are successors of the initial ones. Another interesting finding is that the transaction flows reveal some characteristic behaviors such as long chains, fork merge, and binary tree-like distributions (Ron, Shamir, 2013).

3. Data Management

The data of the Bitcoin transaction graph is publicly available in order to enable the proof-of-work concept for verification of transactions. Sites such as *blockchain.info* or *blockexplorer.com* can be crawled for deriving the entire transaction graph. The data used by our work was collected and to some extent preprocessed by a project of the University of Illinois at Chicago (Brugere, 2013). It contains the time horizon from 01/03/2009 until 04/10/2013. We applied tools developed by Martin Harrigan and Gavin Andresen for extracting data from the *Bitcoin.dat* files in order to construct a user network according to the method introduced by (Reid, Harrigan, 2013). This procedure results in several raw text files (Brugere, 2013). The latest available data for download at the time of writing contained 230,686 blocks with around 37.4 million edges and 6.3 million nodes. The text files were transformed into a specific target format of two tab-separated files, one relationship file and one node file. Once the data had an appropriate structure, it was imported into a relational database. For analyzing the dynamics and topological characteristics of the graph structure, *NetworkX* was used (http://networkx.github.io/) (Hagberg et al., 2008).

4. Analysis Method

In the first step of the analysis several descriptive statistics were calculated. Some of our results were earlier established by Katzenbeisser and Hamacher (2011) and at the Chaos Communication Congress in 2013. Characteristics such as user activity and transaction volume were linked to the Bitcoin exchange rate provided by *Mt.Gox*, which provides services for exchanging Bitcoins (https://www.mtgox.com/).

The second part of the analysis regards the network structure and topology. Since financial transaction networks are always evolving and not static, all measures were applied for different time horizons in order to investigate the dynamics. In the following the network measures are briefly introduced.

The *Degree* distribution captures the structure of networks in terms of the individual connectivity of nodes. The in-degree of a node i is the total number of connections to the node i and is the sum of the ith-column of the adjacency matrix. For the out-degree, the sum of the ith-row of the adjacency matrix is calculated (Gross, Yellen, 2004). One characteristic, often revealed by real networks, is that the degree follows a power law (Clegg, 2006), e.g., as shown by Barabasi, Albert and Jeong (2000) for the World Wide Web and by Inaoka, et al. (2004) in cases of financial transaction networks.

The *Average Clustering Coefficient* measures the global cliquishness on the graph. Watts and Strogatz (1998) applied the clustering coefficient in order to discover the small world phenomenon within several networks. The *Average Shortest Path Length* is defined as the average number of steps along the shortest paths for all possible pairs of nodes and measures the efficiency of information or mass transport in the network (Mao, Zhang, 2013). According to network theory one can determine how efficient Bitcoin is with respect to transactions.

Eigenvector Centrality measures the influence of one node on other nodes. For each node it is defined as the value of the corresponding component of the principal eigenvector of the adjacency matrix defining the network. Accordingly, a node with a high eigenvector score is one that is adjacent to nodes that also have high eigenvector scores (Borgatti, 2005). This measure is essential for discovering central hubs such as exchanges, miners, or “laundry services” that are important nodes in the Bitcoin network.

5. Analysis and Results

The descriptive statistics were applied over the entire time horizon from 01/03/2009 until 04/10/2013. The transaction value per day has a wide range beginning with the initial transaction of 50 BTC up to a daily amount of nearly 30 million BTC (19th September 2012).

Dataset:	3rd January 2009 – 10th April 2013			**Transactions (Relations):** **Economic Entities (Nodes):**			37,450,461 6,336,769
	Median	**Mean**	**Sd**	**Skew**	**Min**	**Max**	**Correl [ExRate]**
Transaction Value [BTC]	173,457	910,053	2,231,647	7	50	29,958,714	0.199
Number of Users	1,637	4,049	5,243	2	1	36,120	0.730
Number of Transactions	3,678	24,084	38,303	2	1	189,284	0.680

Table 1. Statistics of the Bitcoin network.

The distribution of the transaction values is strongly skewed to the left. Another notable result is the high correlation between the number of active users, the number of transactions, and the Mt.Gox exchange rate (BTC/USD), see Figure 1.

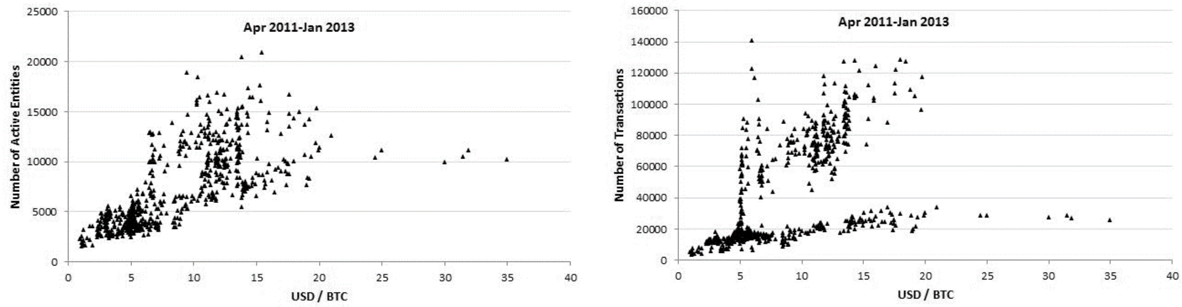

Figure 1. Correlations of user activity (left) and number of transactions (right) to exchange rate.

There are cutoffs at the beginning of trades when the dollar parity was achieved and maintained with negligible changes on 04/13/2011, and at the end when an extremely high exchange rate of around 237 BTC/USD was reached. Both figures show a high heteroscedasticity of the data. This indicates a highly speculative behavior in the network. The relationship will be investigated more thoroughly later on.

Table 2 shows the five largest entities in the network according to their number of public keys. The largest one has the entity ID 11 with over 318,221 public keys. One can also see that this entity is involved in the biggest transactions within the network. All the largest transactions are likely related to each other as indicated by the close time horizon and the entities involved.

Ron and Shamir (2013) conducted an analysis of these Bitcoin flows and came to the conclusion that nearly all major transactions are related. Another interesting result regards the huge amount of tiny transactions. The highest percentage of transactions (6.80%) according to their trade volume corresponds to the transactions of the range from 0.00000001 to 0.00001 BTC. Figure 2 shows the transaction values in a histogram, indicating the peaks of the highest transactions occurring.

Largest Entities		Largest Transactions (Value)				Small Transactions (Value)			
Entity ID	**# PubKeys**	**User_From**	**User_To**	**Date**	**Value**	**Low**	**High**	**#**	**%**
11	318,221	637193	637137	15th Nov 11	500,000.00	0.00000001	0.00001	2546657	6.80
29	209,249	637137	11	16th Nov 11	499,720.70	0.01	0.01099	2187115	5.84
74	109,128	11	636665	16th Nov 11	499,643.98	0.0991	0.10009	777635	2.08
12564	99,939	636665	11	17th Nov 11	499,609.08	…	…	…	…
27	64,993	11	11	17th Nov 11	499,420.95	49.6	50.59	344871	0.92
						10	10.99	332412	0.89

Table 2. Transactions and users in the network.

There is a strong relationship between the exchange rate of Bitcoin and the activity in the network. User activity increases immediately after a peak was reached by the exchange rate. A rolling window was constructed to investigate the relationship for different time windows. The user activity was measured for the last day, last 10 days, last 30 days and the last 100 days. Every rolling window shows a strong relationship but shrinks when extending the time horizon. The correlation coefficient for the last day is 0.736, last 10 days – 0.710, last 30 days – 0.671, and for the last 100 days – 0.641.

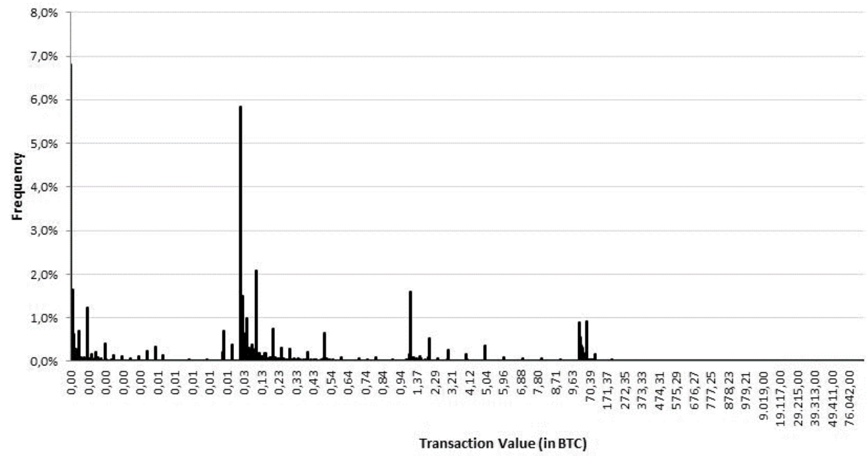

Figure 2. Histogram of the transaction value.

Figure 3 shows the BTC/USD exchange rate provided by the Bitcoin exchange Mt.Gox. There is a cutoff at the end of the time series due to a tremendous increase up to $237. In the following, some events are noted that might explain several of the strong movements in the exchange rate and the respective attention by more potential users of Bitcoin:

a) Start of the public trading of Bitcoins.
b) First time reaching dollar parity on 10th February 2011.
c) Several articles and media attention on Bitcoin, e.g., Forbes, Businessweek, or Bloomberg.
d) Abandonment of Paypal on Cyberlocker sites due to privacy concerns (Dotson, 2012).
e) Cyprus financial system about to collapse, Bitcoin is considered as new safe haven (see Mey, 2013).

Such a strong relationship can also be seen for the number of transactions carried out on the network. This is not surprising since higher activity of users leads to more transactions. The increasing number of transactions follows the exchange rate movements. The correlation coefficient to exchange rate is 0.680 and to user activity 0.928. Between transaction value and exchange rate there is a rather low correlation coefficient with 0.198.

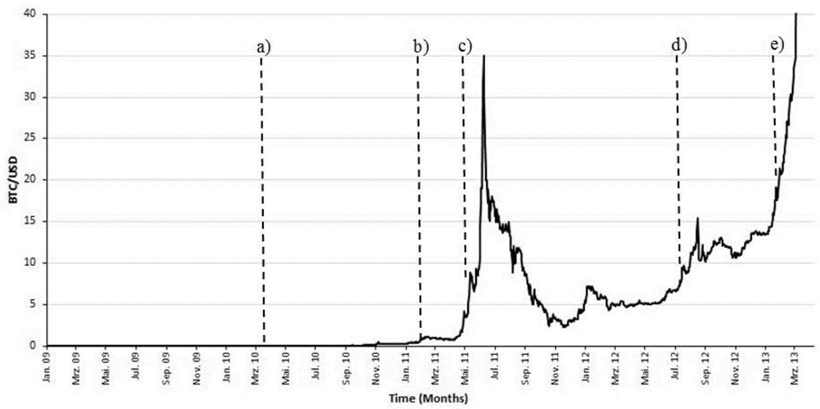

Figure 3. BTC/USD exchange rate and events.

In the following, the focus is placed on network structure. For this, the degree distribution was constructed. For every year since the start of Bitcoin in 2009, the degree k was calculated for every user entity by counting and summing ingoing and outgoing transactions (in- and out-degree). The resulting total degree distribution is drawn on a double logarithmic scale. The distribution also gives an insight into the network usage over time. In the beginning of network activity in 2009, there have been a lot of fluctuations. With increasing network usage the degree distribution seems to converge over time to a scale-free behavior that is also shown by many other real networks. In case of the Bitcoin network this means that the majority of users have a low degree while a small but non-negligible amount of users have a high degree k.

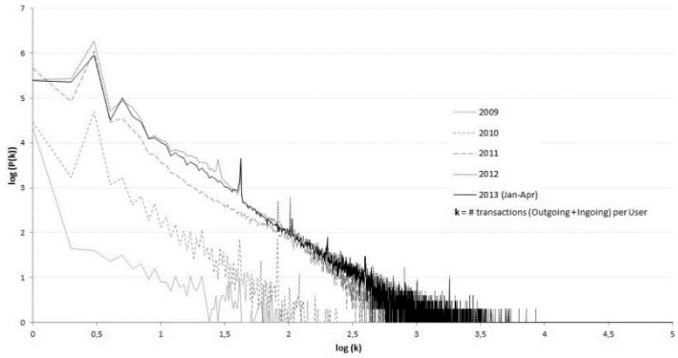

Figure 4. Degree distribution of the Bitcoin network.

Another important metric in the network analysis is the average clustering coefficient. In order to find evidence for a small world network, one can compare the Bitcoin graph to a random network with the same amount of nodes and edges like Watts and Strogatz (1998) did in their analysis. This measure was calculated on a monthly basis for the years 2012 and 2013. For 2011 the calculation was done quarterly. The years 2009 and 2010 were omitted from the analysis due to rather low activity in the network and lots of transactions between the same entities. Over time, the average clustering coefficient is rather high, indicating a small world network.

It can be seen that clustering decreases with increasing activity within the network. In quarter two and three of 2011 the lowest measure was calculated, while the user activity increased in that time period. The same effect can be noted for August 2012 and March 2013. Hence, higher user activity in the network reduces the global cliquishness in the graph.

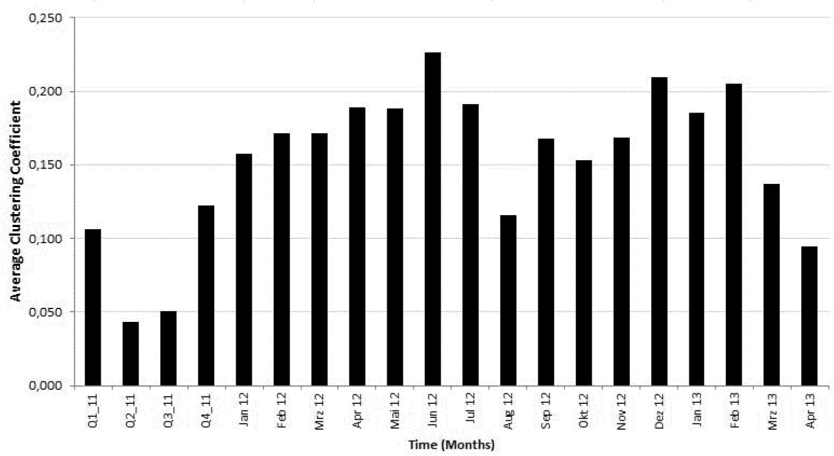

Figure 5. Average clustering coefficient over time.

Due to restrictions on computing power, the metrics average shortest path and the eigenvector centrality are calculated just on the subgraph containing all transactions equal to and higher than 50,000 BTC. Since the average shortest path is only applicable on connected graphs, this metric is calculated for every connected subgraph within the network. In the large transaction network there are 11 disjoint subgraphs. The largest average path length is 125.083 and the lowest is 1.0. The first high value indicates a rather inefficient transfer of Bitcoins through the network according to a common interpretation of this measure. But since users are in control of transferring Bitcoins, this kind of inefficiency might be intended to obfuscate financial transactions.

The eigenvector centrality calculation did not converge to a solution within a reasonable time frame (on convergence see Hagberg et al. (2008)) thus only the degree centrality measure was used. The highest value occurs for the entity 11, which is also confirmed by the visualization of the largest hub in the graph. Degree centrality measures the importance of nodes within a network; the results show that the large transaction network node 11 is the most important node and can be considered as a hub for the others.

Largest Transaction SubGraph			
SubGraphs	**Average Shortest Path Length**	**Entity ID**	**Degree Centrality**
1	125.083	11	0.0769
2	16.142	637193	0.0226
3	2.067	504303	0.00603
4	2.333	675451	0.00603
5	2.000	591334	0.00603
6	1.667	442450	0.00452
7	1.333	2132335	0.00452
8	1.333	…	…
9	1.000	636401	0.00301
10	1.000	…	…
11	1.000	5991405	0.0015

Table 3. Average shortest path length and degree centrality of the largest transaction graph.

6. De-Anonymizing Entities

To demonstrate the possibility of de-anonymizing at least some users in the Bitcoin network, the largest entity in terms of the number of public keys was selected. This entity 11 is also involved in the largest transactions that were carried out on the network. The first approach was to investigate the IP address belonging to the public key that initiates the transaction which is available from the site *blockchain.info*. It needs to be conceded that many IP addresses just reveal information (via Whois) about the last gateway before entering the block chain and thus cannot directly be associated with the real user. But one can receive information on the regional distribution of hosted services and their transactions. Users or business services accepting Bitcoins that are not using hosting services could be uncovered with this approach by using *getaddr.bitnodes.io*.

Another finding is that large and highly active entities providing exchange, laundry, mining, gambling services such as *Mt.Gox*, *SatoshiDice*, or *BTC Guild* are publicly known on the *blockchain.info*, and entity 11 could be identified as the exchange service Mt.Gox. To confirm this result, several transactions until April 10, 2013, in which Mt.Gox was involved were investigated using *blockchain.info* and could be linked to entity 11. Another method is to look up a particular public key of Mt.Gox in the dataset and show that it belongs to the public key pool of entity 11.

7. Conclusion and Outlook

Our research can serve as an exploratory starting point for the application of several techniques, descriptive statistics, and network analysis of the Bitcoin transaction graph. Recent results on the transaction graph were introduced. Standard descriptive statistics and more advanced methods in the field of network analysis were applied.

The results of the descriptive analysis show strongly skewed data series, especially for the transaction value. Another finding is the strong relationship between user activity, transaction volume, and the exchange rate of Bitcoin. One could also see that the largest entity is also involved in the largest transactions carried out in the network, and that the highest amount of transactions is of the smallest possible transaction size. Furthermore, the exchange rate was investigated and related to some events explaining its volatility. A strong relationship of user activity within different time horizons and the exchange rate could be demonstrated.

The network analysis revealed some new findings compared to previous research. We confirmed that the network degree distribution seems to converge to a scale-free network over time. A new contribution was the analysis of the average clustering coefficient, which is an indication for Bitcoin being a small world network as described by Watts and Strogatz (1998). The analysis of the average path length and degree centrality was conducted on a subgraph containing the largest transactions (>= 50,000) in the network. The results show a very large average shortest path of around 125 for the largest connected subgraph, indicating inefficient user-driven transactions possibly aimed at hiding Bitcoin flows. Using the degree centrality measure, the largest hub in the subgraph, entity number 11 (also the largest entity in the entire network), could be found. Future work could aim for de-anonymizing other major hubs of the network possibly by external information and experimental transactions.

The analyses conducted in this work could also be extended by further network measurements. One could investigate the small world character more thoroughly by advanced methods. Also further analysis on centrality can be conducted such as betweenness centrality or current flow betweenness centrality in order to get more insights on important hubs in the network. Further clique and clustering analysis can be used to expose social interaction characteristics of users.

One could also extend the data set with IP address and geo-location data in order to conduct novel research on the geographic characteristics of the network. Then it would be possible to analyze the network structure in different regions and how transactions occur between them. This can lead to a more thorough picture of structures and topology of the Bitcoin transaction graph. All of these analyses can also serve as a starting point in investigating anonymity and economic relationships in Bitcoin on a new structural level.

References

Barabási, A-L; Albert, R; Jeong, H (2000): Scale-free Characteristics of Random Networks: the Topology of the World Wide Web. Physica A: Statistical Mechanics and its Applications 281(1–4): 69–77.

Borgatti, S (2005): Centrality and Network Flows. Social Networks 27(1): 55-71.

Brugere, I (2013): Bitcoin Transaction Network Extraction. https://github.com/ivan-brugere/Bitcoin-Transaction-Network-Extraction. (Access Dec 2013)

Clegg, R (2006): Power Laws in Networks. http://www.richardclegg.org/networks2/SpecialLecture_06.pdf. (Access Dec 2013)

Dotson, K (2012): Paypal's Abandonment of Major Cyberlockers May Become Bitcoin's Big Win. http://siliconangle.com/blog/2012/07/11/paypals-abandonment-of-major-cyberlockers-may-become-bitcoins-big-win/. (Access Dec 2013)

Drainville, D (2012): An Analysis of the Bitcoin Electronic Cash System. https://math.uwaterloo.ca/combinatorics-and-optimization/sites/ca.combinatorics-and-optimization/files/uploads/files/Drainville, Danielle.pdf. (Access Dec 2013)

Hagberg, A; Schult, D; Swart, P (2008): Exploring Network Structure, Dynamics, and Function Using NetworkX. In: Varoquaux, G; Vaught, T; Millman, J (Eds), Proceedings of the 7th Python in Science Conference: 11-15. Project documentation: http://networkx.lanl.gov/.

Inaoka, H; Ninomiya, T; Taniguchi, K; Shimizu, T; Takayasu, H (2004): Fractal Network Derived from Banking Transaction – An Analysis of Network Structures Formed by Financial Institutions. Bank of Japan Working Papers 04(04).

Katzenbeisser, S; Hamacher, K (2011): Bitcoin – An Analysis [28C3]. http://www.youtube.com/watch?v=-FaQNPCqG58. (Access Dec 2013)

Mao, G; Zhang, N (2013): Analysis of Average Shortest-Path Length of Scale-Free Network. Journal of Applied Mathematics, Vol. 2013, Article ID 865643.

Mey, S (2013): Bitcoin Rallye: Hackerwährung als sicherer Hafen für Sparer? http://blog.zdf.de/hyperland/2013/03/bitcoin-rallye-hackerwaehrung-als-sicherer-hafen-fuer-sparer/. (Access Dec 2013)

Nakamoto, S (2009): Bitcoin: A Peer-to-Peer Electronic Cash System. http://bitcoin.org/bitcoin.pdf. (Access Dec 2013)

Gross, J; Yellen, J (Eds) (2004): Handbook of Graph Theory, CRC Press.

Ober, M.; Katzenbeisser, S; Hamacher, K (2013): Structure and Anonymity of the Bitcoin Transaction Graph. Future Internet 5(2): 237-250.

Reid, F; Harrigan, M (2013): An Analysis of Anonymity in the Bitcoin System. In: Security and Privacy in Social Networks, Springer: 197-223. arXiv:1107.4524v2 [physics.soc-ph]

Ron, D; Shamir, A (2013): Quantitative Analysis of the Full Bitcoin Transaction Graph. Lecture Notes in Computer Science 7859: 6-24.

Watts, D; Strogatz, S (1998): Collective Dynamics of Small-World Networks, Nature, Vol. 393, June 1998: 440-442.

ARTICLE 6:

TOPOLOGICAL ANALYSIS OF CLOUD SERVICE CONNECTIVITY[9]

Authors: Benjamin Fabian, Humboldt-Universität zu Berlin, Berlin, Germany

Annika Baumann, Humboldt-Universität zu Berlin, Berlin, Germany

Jessika Lackner, Humboldt-Universität zu Berlin, Berlin, Germany

Journal: Computers & Industrial Engineering 88, pp. 151-165

Publication year: 2015

Abstract

Relying on services in the cloud involves manifold availability risks and concerns. This article focuses on the network reachability of cloud services. We present a study on cloud outages and causes, and analyze the topological connectivity of major cloud service providers (CSPs) by graph-based measures. Our approach is based on the construction and integration of an empirical dataset describing the connections between Autonomous Systems (ASs) of organizations that form the Internet backbone. According to our findings, though the ASs of CSPs generally appear to be better connected than an average AS, they also vastly differ in several connectivity measures, sometimes by more than an order of magnitude. Our results help to identify well-connected CSPs and CSPs that could potentially suffer more from Internet outages, if no additional path redundancy is provided. Our approach can be used by CSPs to assess connectivity beyond their own premises. It can also support cloud service customers during benchmarking and selection of CSPs when high availability is a critical requirement.

***Keywords**: Cloud Computing, Availability, Connectivity, Complex Networks*

[9] This article is provided with kind permission from Elsevier. The original version is available at: https://doi.org/10.1016/j.cie.2015.06.009.

1. Introduction

The market for cloud services (CS) has been growing rapidly in recent years (Panettieri, 2013). Businesses are now capable of saving hardware, software, and maintenance costs by outsourcing IT infrastructure and daily business processes into the cloud. Cloud computing is also transforming classical industry domains such as supply chain management, logistics, and manufacturing (Kong, Fang, Luo, and Huang, 2014; Wu, Greer, Rosen, and Schaefer, 2013). Since CS are independent from devices and operating systems that customers use, they leverage previously unknown flexibility which is further increased by the easy access of CS via the Internet (Mell and Grance, 2011).

However, relying on services in the cloud also involves multiple risks and concerns. Loss of Internet connectivity, anywhere on the path between client and cloud service provider (CSP), could result in inaccessibility of services and data in the cloud, which may result in huge costs (Emerson Network Power, 2011). Therefore, availability and reachability of CS rely on the connectivity of the CSP that can suffer from high traffic load, latency problems, attacks, or disasters. Hence, assessment of CSP connectivity is an important research area that is also of high practical relevance (Bilal, Malik, Khan, and Zomaya, 2014), since several network outages of CSPs and parts of the Internet have been reported in the last few years (see Section 2).

However, most past work focused on assessing and improving the redundancy of data, hardware, and network connections within particular cloud data centers (see Section 3). In contrast, we contribute to cloud availability research by investigating how CSPs are connected to the Internet beyond their premises. This is an important novel approach that can extend existing frameworks for risk assessment as well as CSP selection or certification processes by providing quantitative guidelines.

One of our contributions, we conducted research of publicly known cloud outages over recent years including downtime and the major outage causes. We accumulated data used in this study from search engines and information portals on cloud computing. Next, we showed that network and connectivity problems are an important outage cause. We also analyzed major Internet outages from recent years, which was motivated by an extensive literature review revealing that Internet backbone outages have so far not been given enough attention by cloud research and in practice, given their frequency and high impact.

Our central contribution involves a method for analyzing the topological connectedness of CSPs to the Internet backbone. Our approach is based on modeling Internet connections at the level of Autonomous Systems (ASs), sections of the Internet usually managed by independent organizations (Baumann and Fabian, 2014). We systematically collected AS connectivity data from several public sources and integrated it into a comprehensive graph model. Then, we conducted a topological connectivity analysis using several important graph metrics for connectivity. Based on these results, we compare CSPs to average ASs and rank them according to several connectivity dimensions. We interpret our findings and develop a process of how our method can be used by cloud customers or third-party services to compare CSPs with respect to connectivity.

The rest of the article is structured as follows. First, we survey the CSP outages and major Internet failures over recent years (Section 2). We then discuss the state of research on cloud availability (Section 3), and we present methodology and data for our topological connectivity analysis (Section 4). The results of the study are presented in Section 5 and are discussed in Section 6, which includes limitations, management recommendations, and the CSP selection process based on our method. Section 7 concludes the paper.

2. Cloud and Internet Outages

Many CSPs advertise their service with a promise of 99.9% annual availability, referring to their uptime, for examples of such service level agreements (see Amazon, 2013a, 2013b; Google, 2015; Microsoft, 2014). Though this suggests it to be a reliable service, it can still result in an annual downtime of 8.76 h (Oggerino, 2001). Availability is crucial for businesses and other customers (Lansing et al., 2013) who rely on CS for their daily activities since service downtime can directly result in costs and annoyance. This section will present the results of a systematic study of reported cloud outages over a period of six years.

2.1. Cloud Service Outages

Our search for cloud outages was performed in several stages that finally resulted in a dataset with 159 outages during the period of 2008–2013. This dataset contains information about the year, month, and duration of outages. Additionally, the affected CSP, the search keywords used, the published root cause, and the information source have been recorded. Our search mainly used the Google search engine. Further sources included the websites datacenterknowledge.com (Data Center Knowledge, 2013) where outages are regularly reported, and iwgcr.org (IWGCR, 2013) where outages are listed with downtime and information about whether or not critical data was lost. The latter did not result in discovering additional outages, but was used to validate our dataset. Within our data set, several CSPs stand out by the amount of reported downtime, including Microsoft, Google, Salesforce, Amazon, Rackspace, Research in Motion (RIM), Apple, Dropbox, and Hosting.com.

2.2. Analysis of Downtime

The accumulated downtime for all relevant services is shown in Table 1. This table provides the downtime (in days) for each year of the observation period (2008–2013) accumulated across all CSPs with reported outages. In addition, the average, median, and maximum and minimum amounts of downtime are displayed for each single year as well as for the dataset in total. In some instances, the reported downtime varies between sources. In this case, the most frequently reported duration was chosen.

Year	Σ Downtime	Max	Min	Median	Mean
2013	17.01	3	0.008	0.125	0.364
2012	16.59	2	0.008	0.167	0.448
2011	18.10	3	0.019	0.250	0.953
2010	10.70	3	0.042	0.083	0.508
2009	17.11	6	0.042	0.208	1.207
2008	7.03	1.25	0.031	0.250	0.414

Table 1. Accumulated CSP downtime per year (in days).

2.3. Cloud Outages Related to Networks

Outages that are directly related to the network connection of CSPs occurred quite frequently. The prevalence of this type of failure, along with its percentage of 17.43% of overall downtime, explains its relevance and motivates the need for further studies. Several causes were stated as reasons for networks to fail. The most frequent causes involved capacity and load issues, misconfigured network devices, problems with the network infrastructure, faulty nodes and core switches along with Internet Protocol (IP) layer and Domain Name System (DNS) problems. Table 2 shows a selection of such events identified by our study.

Considering the overall downtime caused by network outages, RIM reportedly suffered most with 5.75 days over the period under observation. Microsoft follows with a total of 4.75 days, then Salesforce with 2.44 days of downtime. Also with respect to the maximum downtime caused by a single-event network outage, RIM leads with three days, followed by Microsoft and Salesforce with 2.6 and 1.8 days respectively.

Figure 1 shows the percentage of downtime caused by network outages for a certain CSP in percentage of its total amount of downtime. According to the possibly limited view of our data, RIM was the most vulnerable CSP to such kinds of outages, followed by Salesforce, Apple, Microsoft, Amazon, Rackspace, and Google.

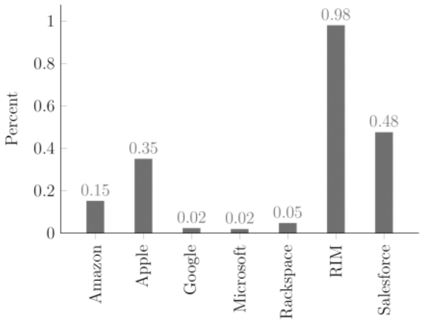

Figure 1. Ratio of network-related to total downtime.

Clearly, there are some limitations to our research. First of all, even though due care was applied during the search, the public data gathered could be incomplete. Moreover, some outages may not have been reported publicly. Furthermore, it is noteworthy that the outage causes we applied are the officially stated causes.

2.4. Large Internet Outages

Our statistics of network outages contain many examples of local problems at the CSP. But outages may also be caused by a whole range of similar failures at upstream Internet Service Providers (ISPs). A single network failure does not necessarily result in a full service outage if the CSP has several connectivity paths to choose from. Regarding the Internet as a graph, full network outages can also be the consequence of a vertex that is not well connected to the entire network topology.

Internet failures and attacks are not just theoretical assumptions. A quantitative overview of operation failures in the IP-based backbone is given in Markopoulou et al. (2008). But Internet availability does not only depend on network infrastructure and functionality but also on several additional factors such as the power grid. Some large-scale disasters and intentional attacks have already happened in reality (Baumann and Fabian, 2014; Sterbenz et al., 2010; Wu et al., 2007). An example of one such natural disaster is Hurricane Katrina that occurred in 2005 (Renesys, 2005). Some of the most damaged areas were Louisiana, Mississippi, Alabama, Florida, and Georgia, resulting in a fraction of between 8% (in Louisiana) and 38% (in Mississippi) of the locally situated networks becoming unavailable (Renesys, 2005). A second example is the massive Taiwan earthquake of 2006 which damaged seven undersea cables and rendered up to 4000 networks unavailable or not working properly (Renesys, 2007).

Errors, both human and system, can also have severe impact. A series of aligned malfunctions caused a blackout in Canada and the Northeast USA in 2003 which affected Internet access significantly (Renesys, 2003). A maximum of around 3175 networks were unavailable due to the outage, affecting over 50% of ASs in this area and more than 1700 organizations (Renesys, 2003). Other errors include accidental cable cuts, which seem to happen regularly such as the cut to the Mediterranean Sea Cable in

2008, and different forms of misconfigurations, such as the one by the Pakistan Telecom in 2008 (Sterbenz et al., 2010, p. 6).

Intentional attacks are another challenge including (distributed) DoS, Internet worms, or acts of terrorism. Famous examples of worms include Nimda and Code Red v2, which both emerged in 2001 (Renesys, 2002). Apart from the destruction they caused to computers, these worms resulted in enormous Border Gateway Protocol (BGP) update rates leading to routing instabilities (Renesys, 2002). Even if the Internet in its entirety is apparently quite robust (Sterbenz et al., 2010, p.7) against random failures (Baumann and Fabian, 2014), such disruptions still tend to massively affect Internet connection in specific areas.

Year	**Month**	**Provider**	**Outage cause**	**Outage category**	**Minutes**
2008	February	Amazon S3	Network capacity	External cause	180
2008	February	Research in Motion	Network infrastructure problem	Human error	240
2008	February	Salesforce	Slowdown in storage sub-systems	External cause	360
2008	August	Google Gmail	Outage in contacts system	Software failure	90
2009	January	Salesforce.com	Network device failed	Hardware failure	120
2009	September	Google Gmail	Network capacity	External cause	100
2009	December	Rackspace.com	Router configuration error	Human error	60
2010	January	Microsoft BPOS	Problem with network infrastructure	Hardware failure	90
2010	January	Salesforce.com	Failure of a network device	Software failure	60
2010	November	PayPal	Failed shifting of traffic to another data center	Hardware failure	90
2011	January	Microsoft Windows Live	Load balancing issues	Software failure	240
2011	April	Amazon Web Services	Network configuration error	Human error	780
2011	May	Microsoft BPOS	Malformed e-mail traffic, DNS failure	External cause	120
2011	June	Microsoft BPOS	Network equipment issues	Hardware failure	120
2011	September	Microsoft Windows Live	Failed update of network traffic balancer	Software failure	3720
2011	November	Apple iPhone 4s Siri	Network outage	Hardware failure	1440
2012	July	Windows Azure	Misconfigured network device	Human error	150
2012	August	Windows Azure	Failure of network outage prevention	Software failure	144
2012	September	GoDaddy	Corrupted router data tables	Software failure	360
2012	October	Google App Engine	Network capacity	External cause	240
2012	December	Microsoft Xbox Live & Azure	Faulty nodes	Hardware failure	2160
2013	January	Research in Motion	Provider router problem	External cause	4320
2013	February	Microsoft 365 & outlook.com	Network configuration failure	Human error	120
2013	May	Rackspace	Network capacity	External cause	45
2013	May	Rackspace	Network capacity	External cause	180
2013	July	Salesforce	DNS fault	Software failure	2400
2013	July	Rackspace	Network capacity	External cause	60
2013	July	Rackspace	IP Issue	Software failure	12

Table 2. Selected network-related CSP outages (2008–2013).

3. Related Work on Cloud Availability

Availability and reliability are important requirements for any online business service and need to be addressed through proper planning and systematic mathematical modeling (Kryvinska et al., 2010, 2011; Kryvinska and Strauss, 2013). In particular, customers of CS critically depend on service availability (Agapi et al., 2011; Ermakova et al., 2013) and should consider related risks before any cloud-sourcing decision is made. Corresponding assessments should be part of any process for ranking and selecting CSPs (Garg et al., 2013; Hajjat et al., 2010; Hu et al., 2014; Li et al., 2010; Toosi et al., 2014) and formal certification processes (Lansing et al., 2013). However, even though many articles and security guidelines for cloud computing (CC) discuss service and data availability in the cloud (Catteddu and Hogben, 2009; Subashini and Kavitha, 2011), they usually focus on hardware and data redundancy at the CSP site while only briefly discussing general Internet backbone outages.

Some research has been conducted on actively probing the availability of selected CSPs: In Li et al. (2010), the authors present *CloudCmp*, a CSP comparison tool that includes measurement of network throughput and path latency. This tool is used for comparing four major CSPs. An assessment of methods for measuring end-to-end availability of CSPs is presented in Hu et al. (2014), including Internet Control Message Protocol (ICMP) traceroutes and Hypertext Transfer Protocol (HTTP) requests. These methods are evaluated by measurements for three CSPs. Another interesting and complementary direction for determining CSP availability is based on mining social media, such as Twitter, for outage reports (Motoyama et al., 2010).

More closely related to our work, Bilal et al. (2013) analyzes reliability with a focus on a single cloud data center. Their approach is also based on graph theory which they apply to compare different network architectures that a CSP can implement for designing a data center network. In Gill, Jain, and Nagappan (2011), a detailed study of network failures in data centers is conducted. DCell, a novel network structure for networking within a single data center, is presented in Guo et al. (2008). Similarly, Manzano, Bilal, Calle, and Khan (2013) discusses internal connectivity of data center networks, focusing on the comparison of three important architectures (ThreeTier, FatTree, and DCell). In their study, the authors use artificially generated topology instances and two global graph metrics (A2TR and μ-A2TR), i.e., metrics that quantify the connectivity of the entire network. In contrast, we are focusing on AS-level connectivity of CSPs based on empirical connectivity data and adopt several local graph metrics that quantify connectivity of CSPs regarded as single vantage points.

A formal approach for modeling and assessing cloud availability is presented in Frattini, Trivedi, Longo, Russo, and Ghosh (2014). Mitigation attempts for increasing availability include geo-aware replication (Wu et al., 2013) and general proposals for cloud federation, intercloud, or multi-cloud approaches such as (Abu-Libdeh et al., 2010), for example, by distributing data to multiple CSPs in a redundant fashion. Such redundancy could be achieved by cryptographic methods that could also increase data confidentiality at the same time (Ermakova and Fabian, 2013; Fabian et al., 2015).

Our topological analysis is complementary to all of these approaches and investigates CSP connectivity beyond their local premises by investigating the “deep” Internet. Though several projects have tried to map the Internet at the level of individual routers and their links at particular points of time as well as by active ICMP probing (Cunha et al., 2011; Katz-Bassett et al., 2008; Quan et al., 2013), a sufficient map or graph at this level of resolution is still not available and would also become outdated very quickly due to highly dynamic changes. In contrast, Internet maps at the level of ASs are mostly based on BGP routing announcements. Though still complex, they are easier to construct and analyze while maintaining essential information on connectivity.

Research on Internet resilience based on graph metrics is an established but still evolving field. Several graph models have been used to imitate the Internet structure. Classical network modeling applies the Erdös and Renyi (1959) and Newman (2003) graph model or the scale-free model developed by Barabasi and Albert (Newman, 2003). Early articles such as (Albert et al., 2000; Barabasi and Albert, 1999) study attack and failure tolerance of complex networks and particularly the Internet at the AS level, using the diameter as a global connectivity metric. In Dolev et al. (2006), the Internet is also studied at the AS level with additional restrictions caused by policy-driven routing. Xiao et al. (2008) develops global attack procedures that are based only on local information. In Schneider et al. (2011), the authors focus on malicious attacks and develop a theoretical method that makes the AS-level network more robust by interchanging edges of node pairs. From a more general perspective, Sterbenz et al. (2010) and Smith et al. (2011) present resilience frameworks and metrics. Surveys of robustness metrics are given in Mahadevan et al. (2006), Manzano et al. (2011), and Manzano et al. (2012). In Deng et al. (2011), the authors analyze the k-fault tolerance of the Internet on the AS level, defined as the reachability of a pair of nodes in the network after the removal of k nodes. In Neumayer and Modiano (2010), network reliability is studied with respect to geometric correlation and global metrics.

Though several analyses of the global AS-level Internet and its connectivity properties have been conducted, no one has yet to our knowledge taken the step towards local and individual risk analysis for selected ASs, such as those used for CS. Global AS graph analyses have been focusing on stability and communication ability of the Internet in its entirety in cases of random errors or targeted attacks on important vertices or edges (Baumann and Fabian, 2014). In contrast, a local connectivity analysis for ASs of CSPs has not been conducted before, to the best of our knowledge.

4. Methodology

This section introduces concepts and methods of our analysis. First, a basic introduction to the current Internet topology is provided, followed by a presentation of relevant graph metrics. Finally, supporting measurements for conducting the study are discussed.

4.1. Topology of the Internet

The Internet can be understood as a large network of ASs (Dolev et al., 2006). Each AS encompasses a network of hosts with Internet Protocol (IP) addresses and communicates with other ASs via border routers (Dimitropoulos et al., 2007) and an exterior routing protocol (Dolev et al., 2006), usually the BGP. Border routers exchange routing information (BGP announcements) about the reachability of ASs along with lists of feasible AS paths to other networks (van Beijnum, 2002). A globally unique Autonomous System Number (ASN) (ARIN Board of Trustees, 2013) identifies each AS (Baumann and Fabian, 2014). BGP supports the decision of choosing a path, subject to performance and business constraints, for a particular communication (Dolev et al., 2006).

The AS-level graph used in this article is based on consolidated data from the end of the year 2012. We used three main sources containing information about ASs and their relationships to develop this Internet graph: BGP routing tables, traceroute data, and the Internet Routing Registry (IRR). In the case of the BGP routing tables, we combined data of two publicly available projects. The first is the *Center for Applied Internet Data Analysis*(CAIDA) *AS Rank* project that mainly uses data derived from Oregon RouteViews (University of Oregon, 2013) and RIPE-RIS (RIPE, 2013). The second source is the dataset provided by the University of California in Los Angeles (UCLA) (UCLA, 2013) consisting of data also derived from RouteViews (University of Oregon, 2013) and RIPE-RIS (RIPE, 2013) as well as from route servers and BGP looking glasses.

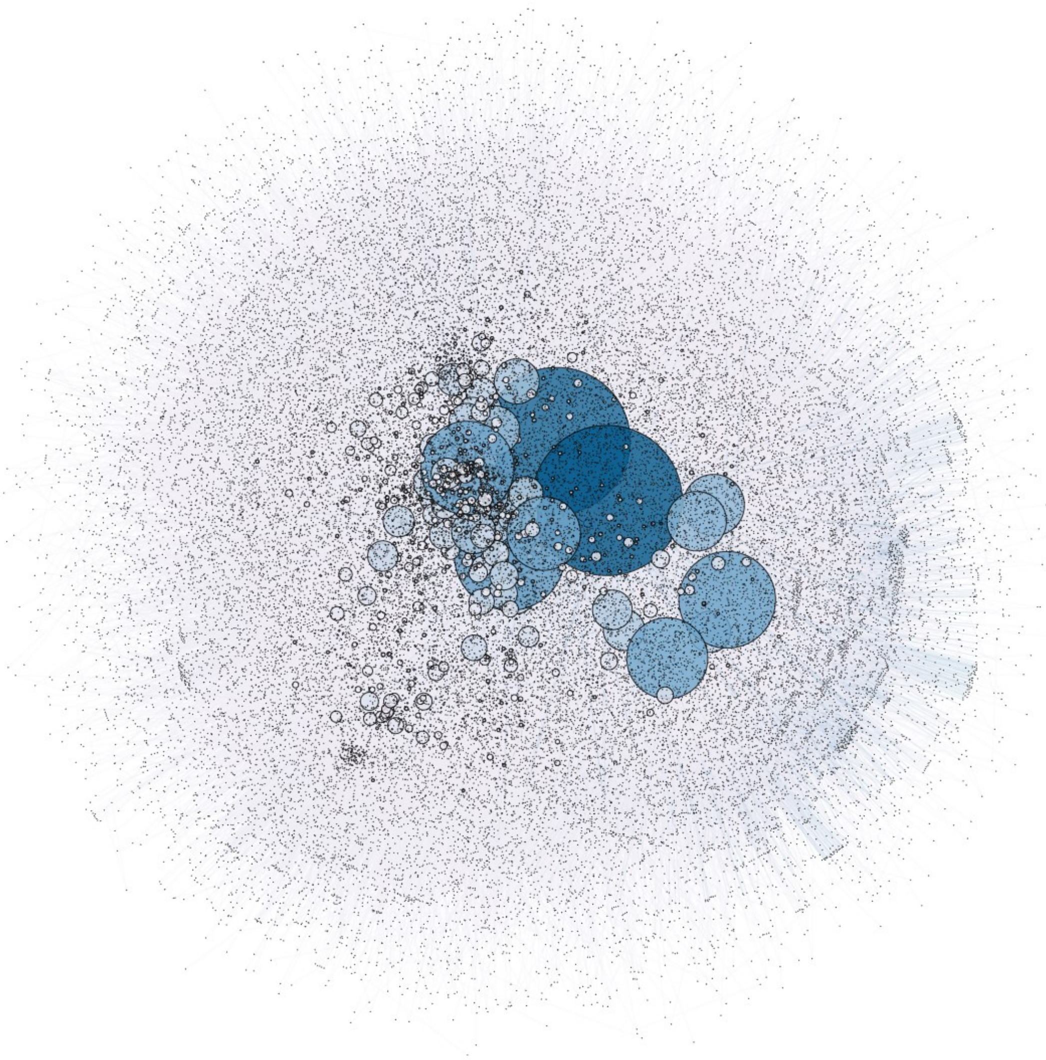

Figure 2. Full AS graph.

The traceroute data is based on the *Macroscopic Topology Project* of CAIDA that is centered around the *Archipelago* (Ark) measurement tool (CAIDA, 2013). With the help of several monitors located around the world, this project collects IPv4-routed address paths. Traceroute methods discover only IP paths, which then need to be assigned to the associated ASs, resulting in an AS connection dataset. Finally, we also used IRR data by collecting the data files of all available 34 IRRs (Merit Network Inc., 2013). Only those AS paths which were not changed later than in 2012 as well as mentioned by both participating ASs were considered as relevant.

All data were cleaned up, converted, and integrated to a common graph storage format. This procedure resulted in a graph consisting of 44,397 vertices that are connected by 199,073 edges. Each vertex represents a particular AS and is labeled with its respective ASN. The edges of this undirected graph represent communication links between ASs (Figure 2). An excerpt of this graph, displaying only the direct connections between selected CSP ASs, is shown in Figure 3. Larger nodes feature a higher degree, i.e., they have more direct neighbors.

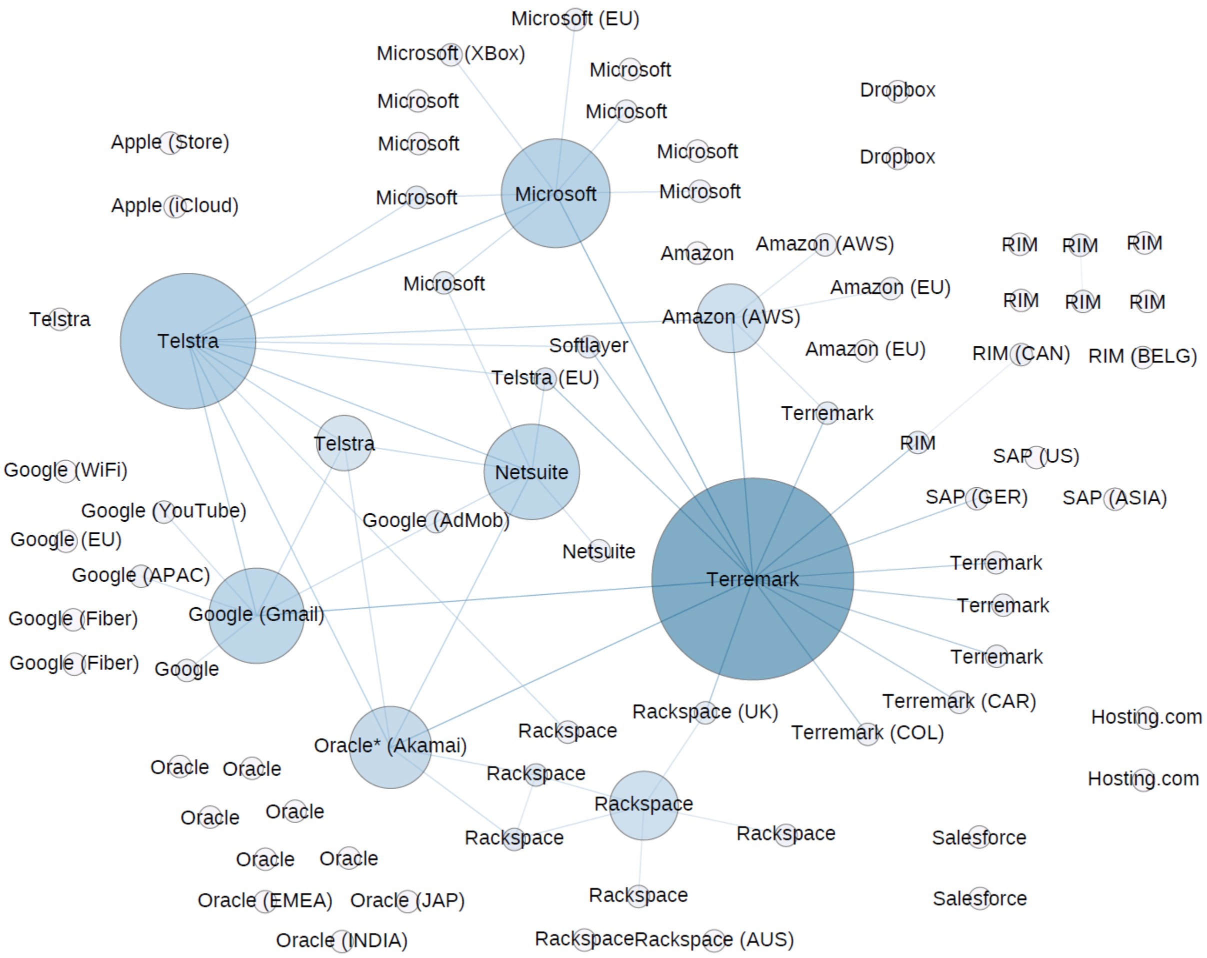

Figure 3. *Subgraph: direct connections between CSP ASs.*

4.2. Graph Metrics

Our article focuses on local connectivity of vertices representing CSPs rather than global properties of the Internet graph. This means that the focus is placed on connectivity of a certain vertex, representing a particular AS such as a CSP, and its neighborhood in the graph. If a vertex is well connected (i.e., has a high connectivity), it can be available even in the case of partial network outages. Such a vertex should have various path options to choose from even if other vertices, through which data originally would have been transmitted, should fail. Clearly, a high number of neighbors is an indicator for a vertex being well connected. Furthermore, a one-hop neighborhood (i.e., the set of neighboring vertices reachable within one hop) is itself well connected and also increases the connectivity of a vertex. The vertices with the shortest distances are located in the topological *center* of a graph. A close distance of a vertex to the center of the network also indicates well-connectedness. Therefore, data is able to spread rapidly. The closer to the center a vertex is located, the better its connectedness. In order to capture these intuitions more formally, we apply a set of graph measures that correspond to local properties of a node rather than focusing on global graph properties. For our newly developed analysis programs we utilized the Python library *NetworkX* (Hagberg et al., 2008) which supports the study of the structure, dynamics, and functions of complex networks (NetworkX, 2014).

4.2.1. Distance Measures

First of all, one should investigate how topologically close a vertex is located with respect to the rest of the network. Distance measures determine the amount of hops between a source vertex and a target vertex. They are also an indicator for network flow and information spread (Boccaletti et al., 2006). There are two important local measures for distance with a perspective of a single vertex. The first is the *eccentricity (ECC)* of a vertex v, which is defined as the maximum distance from vertex v to all other vertices in the graph. The larger this value, the longer the path between the vertex to the topological center of the graph, which is formed by several vertices with minimum ECC (Santhakumaran, 2010).

The second measure is the *Single Source Shortest Path Length (SSSPL)* which refers to the average of the minimum distances $d(v, j)$ between a vertex v and all other vertices j of the graph. Here, "length" always refers to the number of hops between source and target vertices (NetworkX, 2014). SSSPL therefore provides the average amount of hops that are needed for fast communication from the source vertex to all other vertices in the graph. The shorter the average path between a vertex and all other vertices in the graph, the less its path-related risks of failure, and the more centrally the vertex is located resulting in more path options in case of connection failures.

4.2.2. Centrality Measures

Centrality measures provide information about the relative importance and connectivity of a vertex in the graph. Four of the most frequently used centrality measures are degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality. The *degree centrality* $(DC) d_v$ of a vertex v counts the number of edges that the vertex is connected to Boccaletti et al. (2006). It is a fundamental measure for the direct connectedness of a vertex. The more edges a vertex has, the more paths are available and the more alternatives exist in case one edge or neighbor is disconnected. In other words, the higher the degree of a vertex, the better its connectivity to the network (Mahadevan et al., 2006, pp. 19–20). The terms 'DC' and 'degree' are used interchangeably in this article. The degree measure reflects that connectivity of a vertex grows with the number of the vertices it is directly connected to. But it only considers one-hop-relationships of a vertex, neglecting the indirect impact of how well neighboring vertices are connected deeper in the graph.

Eigenvector centrality (EC) extends the view of DC (Bonacich, 1972). EC measures the importance of a vertex but also takes into account that different connections from one vertex to others are not equally important. EC favors connections to vertices that are central and therefore more relevant for connectivity. For a vertex v, EC is calculated as follows (Bonacich and Lloyd, 2001):

$$EC_v = \frac{1}{\lambda_{max}(A)} \sum_{j=1}^{N} a_{vj} x_j \tag{1}$$

The total number of vertices is represented by N; $A = (a_{vj})$ is the adjacency matrix and $x = (x_1, \dots, x_N)^T$ refers to an eigenvector for the maximum eigenvalue $\lambda_{max}(A)$ of A. Vertices with a high EC are connected to vertices that respectively also have high EC scores. In general, vertices that are well connected to highly central nodes can be considered better connected. Since the first connection to such a central node could still be just a single link, EC and degree formalize different aspects of connectivity.

Betweenness centrality (BC) of a vertex v, often also referred to as *load* of v (Boccaletti et al., 2006), measures the number of shortest paths passing through a vertex (Mahadevan et al., 2006). BC can be considered as the expected traffic load on this vertex. This heuristic is, however, somewhat limited as it assumes the traffic to be evenly distributed over the whole network (Mahadevan et al., 2006, p. 23; Freeman, 1977). But in the absence of complete data on traffic flows, this metric can serve as another relevant connectivity indicator. BC is computed as follows:

$$BC_v = \sum_{j,k} \frac{n_{j,k}(v)}{n_{j,k}} \tag{2}$$

The number of shortest paths connecting vertices j and k that are passing through vertex v is represented by $n_{j,k}(v)$ while $n_{j,k}$ represents the total number of shortest paths connecting j and k. BC is usually normalized by $N(N-1)$ as this is the maximum possible BC score of a vertex in a network of fixed size.

Closeness centrality (CLC) measures the closeness of a vertex to all other vertices. When calculating CLC, literature frequently refers to the equation for *Standardized CLC* of a vertex v, which is computed as the inverse of the sum of the shortest path distances from v to all other $N-1$ vertices (Okamoto et al., 2008). The higher the CLC of a vertex, the shorter the distance between this vertex and all others, which therefore indicates greater centrality and connectivity. In the equation, $d(u,v)$ is the shortest-path distance between v and u.

$$CLC(v) = \frac{N-1}{\sum_{u=1}^{N} d(u,v)} \tag{3}$$

Local average neighbor degree (LAND) is another metric that provides information on alternative path options for a vertex and takes the degree of its neighbors into account (Mahadevan et al., 2006). In an AS context, this metric shows if an "AS of a given degree preferentially connects to high or low degree ASes" (Mahadevan et al., 2006, p. 26). A higher LAND score is an indicator for a better connected vertex. LAND focuses on a single vertex and is calculated as follows, where $U(v)$ is the set of direct neighbors of vertex v and k_j is the degree of a vertex j which belongs to $U(v)$:

$$LAND_v = \frac{1}{|U(v)|} \sum_{j \epsilon U(v)} k_j \tag{4}$$

Local node connectivity (LNC) is the final metric utilized in this article. It does not refer to a single vertex but to a pair of nodes. LNC measures the minimum number of vertices that must be removed with all of their edges in order to disconnect two vertices, i.e., to destroy all paths between them (Kammer and Taubig, 2005, p. 142; NetworkX, 2014).

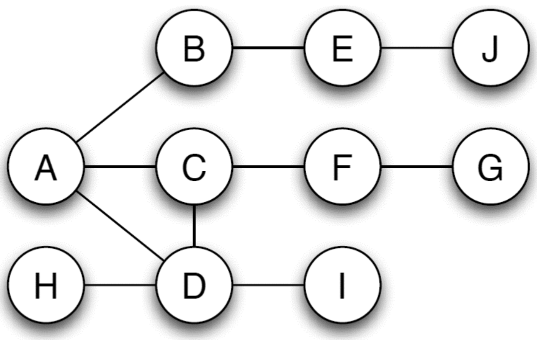

Figure 4. Example graph for illustration of metrics.

4.2.3. Metrics Example

For an illustration of the discussed metrics, assume an example graph as presented in Figure 4. The corresponding metrics are given in Table 3. As the results show, it is clearly possible to identify weakly connected nodes such as G,H,I, and *J* by using these metrics. Moreover, well-connected and important nodes such as A,C and *D* can also be easily differentiated from average nodes. Those well-connected nodes tend to have a higher DC, a shorter SSSPL and lower ECC values than the other nodes. Furthermore, the EC and CLC scores assume higher values, indicating that nodes such as A,C and *D* are more central with respect to the structure of the whole network.

Node	ECC	SSSPL	Degree	EC	BC	CLC	LAND
A	3	1.7	3	0.49	0.50	0.53	3.0
B	4	2.1	2	0.23	0.39	0.43	2.5
C	4	1.8	3	0.49	0.39	0.50	3.0
D	4	1.8	4	0.55	0.42	0.50	2.0
E	5	2.7	2	0.11	0.22	0.33	1.5
F	5	2.4	2	0.22	0.22	0.38	2.0
G	6	3.2	1	0.09	0.00	0.28	2.0
H	5	2.6	1	0.21	0.00	0.35	4.0
I	5	2.6	1	0.21	0.00	0.35	4.0
J	6	3.5	1	0.04	0.00	0.26	2.0

Table 3. Metrics for example graph.

4.3. IP-level Traceroute Measurements

Traceroute is a popular procedure for determining paths from a host to a certain destination in an IP network (Microsoft Technet, 2013). It is used in this article for investigating IP-level paths and connectivity of selected CSPs by using the recorded IP addresses for identifying and validating the traversed ASs. In order to discover a certain path, the ICMP-based traceroute variant sends successive Internet Control Message Protocol (ICMP) Echo Request messages to the destination while successively increasing the Time to Live (TTL) field of the IP header with each attempt (Microsoft Technet, 2005). The tracing process is started by sending an Echo Request with a TTL of one to the destination. Each router on the path that receives the request is required to reduce the TTL by one before forwarding it. If the TTL equals zero, an error message is sent back to the source that then increases the starting TTL by one and sends another Echo Request. Traceroute stops either if a maximum amount of hops is reached or if a response from the target destination is received (Microsoft Technet, 2013). Some router configurations, however, prevent answering such requests. The tool records the IP address of each router and the time it took for this hop to answer.

5. Analysis of Cloud Connectivity

5.1. Identification of Cloud Autonomous Systems

In the following, the AS-level graph described in Section 4.1 is analyzed in order to investigate the connectedness of major CSPs. We began by selecting the ones that were known to have suffered some downtime in recent years (see Section 2). Five additional providers (Oracle, SAP, Softlayer, Terremark, and Netsuite), which have not been publicly noted for experiencing outages thus far but are part of the top ten CSPs of the Talkin'Cloud 100 ranking, have been included in our investigation as well as the Australian CSP Telstra because of its location and three well connected ASs. The final set of 15 CSPs that we investigated in depth is listed in Fig. 5. It was first necessary to identify the ASNs used by the CSPs in order to apply the metrics introduced in Section 4.2 to the corresponding vertices. For this purpose, several data sources from the Web were consulted. These sources provide a catalog of ASNs with their corresponding organization. One list that maps ASNs to organizations provided our first insight about CSP ASs (BGP Reports, 2013). ASNs found on this list were cross-checked and augmented with the help of two websites that provide an ASN-INFO function (Turing Technology Services, 2013; Neustar Inc., 2013). For almost every CSP, it was possible to determine a best-connected primary AS (in terms of the important DC metric) and to verify that the primary AS is being used for most of their services (except for Telstra). This was verified with the help of unique host names of the CS, e.g., aws.amazon.com as well as WHOIS and DNS lookup services. Additionally, a web browser add-on that displays the ASN of every visited website was applied. This add-on extracts its information from default-free zone BGP data as well as from WHOIS entries of regional Internet registries (Internet Business Solutions AG, 2013).

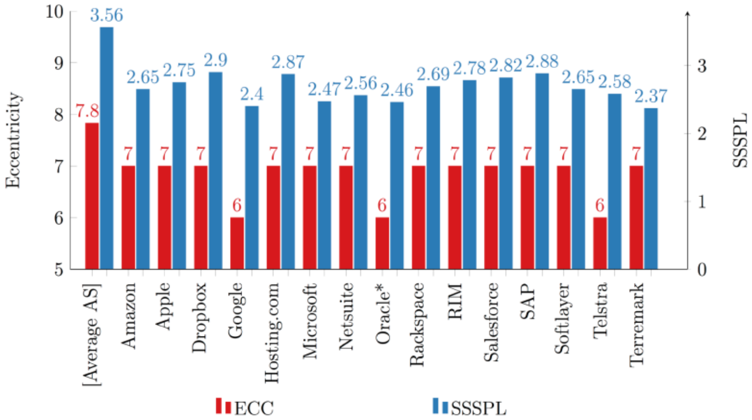

Figure 5. Distance measures for all CSPs.

The focus on the primary AS for each CSP in the following analysis is based on the rationale of providing a realistic best-case scenario for connectivity. We chose not to combine connectivity for all ASs of a single CSP into a single score since non-primary ASs are often located in different regions, are much less connected (except for Telstra), or serve special purposes. In general, they cannot easily replace running services of the primary AS if it should fail since many configurations, data transfers, and routing changes would be necessary. Furthermore, some special cases need to be mentioned. Some major CSPs, such as Amazon, Google, and Microsoft, have several data centers around the globe which are using different ASs. In all of these cases, the main AS in the US is much better connected than the other regions and was therefore chosen as the primary AS.

For other regions and ASs, Table 8 in Appendix A can be consulted. Moreover, we found that CSPs can influence one another, such as the case of Dropbox and Oracle providing offerings based on Amazon's CS. Since such interactions between CSPs can be complex and could make use of important services of the "front-end" company (e.g., authentication), we included their own primary ASs into our analysis. In the case of Oracle, the AS for the website cloud.oracle.com is globally served by Akamai, based on a long-standing business relationship for distributing software as a service (Akamai, 2001). This AS is much better connected than all of Oracle's other ASs, and we decided to use it as a best-case scenario for services of this CSP (noted as Oracle∗ in all figures and tables).

5.2. Connectivity Results

The average results for all vertices of the global AS-level graph are presented in Table 4. They will serve as a baseline for comparison.

Metric	Result
Eccentricity (ECC)	7.83
Single-Source Shortest Path Length (SSSPL)	3.56
Degree Centrality (DC)	8.97
Eigenvector Centrality (EC)	0.00096
Betweenness Centrality (BC)	0.000058
Closeness Centrality (CLC)	0.29
Local Average Neighbor Degree (LAND)	703.28

Table 4. Average results for all ASs (mean values).

5.2.1. Distance Measures

5.2.1.1. Eccentricity

ECC for most vertices under observation is either seven or eight, while the lowest ECC is six (Google, Oracle, and Telstra). In all cases, the primary AS has a lower ECC value than an average vertex (7.8302) of the whole Internet graph and is therefore located closer to the center of the network. The ECC for the entire AS-level graph varies between six and eleven. Santhakumaran (2010) stated that the center of a graph is formed by several vertices with minimum ECC. For the center, the exact ECC threshold can be a matter of debate. However, Google, Oracle, and Telstra are CSPs very close to the center or are even part of it along with other vertices.

5.2.1.2. Shortest Path Length

The SSSPL between a fixed CSP AS and all others in the graph ranges from 2.3 to 4.08 (both Oracle). The lower this value is, the more central the vertex in the network is. The SSSPL for the primary ASs is below three in all cases. With respect to both ECC and SSSPL, all of the 78 analyzed vertices (including the non-primary CSP AS, see Table 8) are clearly closer to the center than an average AS.

5.2.2. Centrality Measures

5.2.2.1. Degree

The average DC in the AS graph under observation is nearly nine (8.968). Even though the highest degree of an AS is 4330, more than half of the ASs (i.e., 27,156) have a very low degree of either one or two. The primary ASs were found to be significantly better connected in terms of degree than an average vertex in the graph. However, the additional ASs (Table 8) usually have a significantly lower degree than the primary AS, often even lower than an average AS. Fig. 6 shows the degree of each primary AS.

The best-connected CSP vertex in this observation is AS 23148 (Terremark) with a degree of 897. Terremark is also represented by three non-primary ASs which have a low degree compared to the degree of an average vertex (Table 8). Netsuite's primary AS is connected to the rest of the AS graph by 486 edges. It has only one additional vertex with a degree of three. Similarly, Salesforce, Apple, Dropbox, and Hosting.com are also represented by only two ASs in the graph.

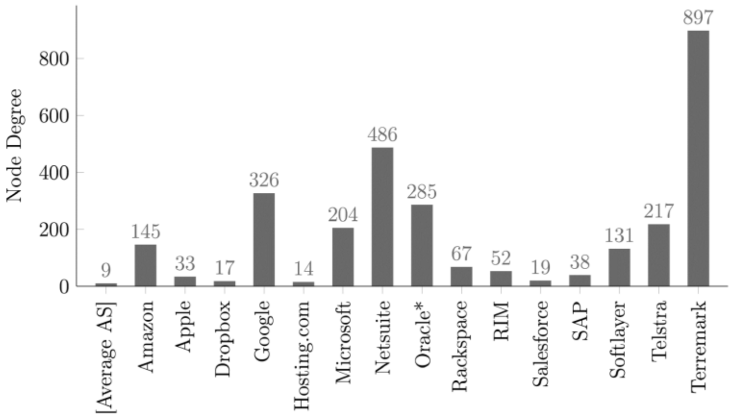

Figure 6. Degree.

5.2.2.2. Eigenvector Centrality

Vertices with a high EC score are connected to other central vertices, which tend to make them more robust in case of network failures (Landherr et al., 2010). EC values for primary ASs are shown in Fig. 7. All the values are above average, but for some CSPs only slightly so. Terremark has the highest EC score (0.099) followed by Oracle and Google. For all CSPs that are represented with one additional AS such as Apple, Dropbox, Hosting.com, and Netsuite, this additional vertex has a low EC. The additional ASs of Oracle, Terremark, and SAP are also connected to less important neighbors. Roughly 45% of all analyzed vertices have an EC greater than an average vertex.

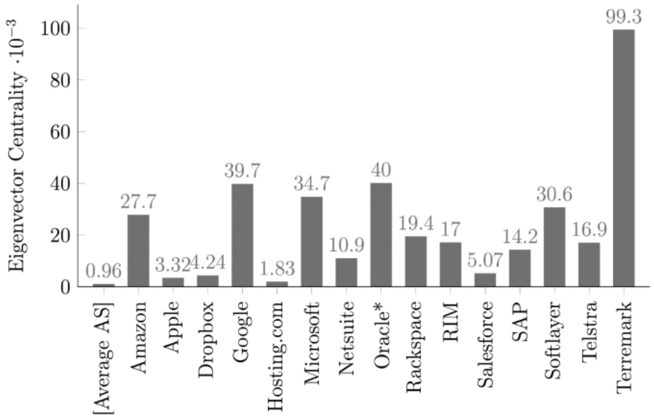

Figure 7. *Eigenvector Centrality.*

5.2.2.3. Betweenness Centrality

BC scores for all primary ASs are shown in Fig. 8. For every company, the primary AS has the highest BC. The most important vertices of the dataset in terms of BC are AS 15169 of Google (0.0076) and AS 3561 of Netsuite (0.0074). For 32 investigated vertices, the BC score is zero which means that no shortest path is leading through them, indicating that those ASs are not important as transit AS for the communication between others. This could be relevant for cloud-based data sharing between globally distributed clients.

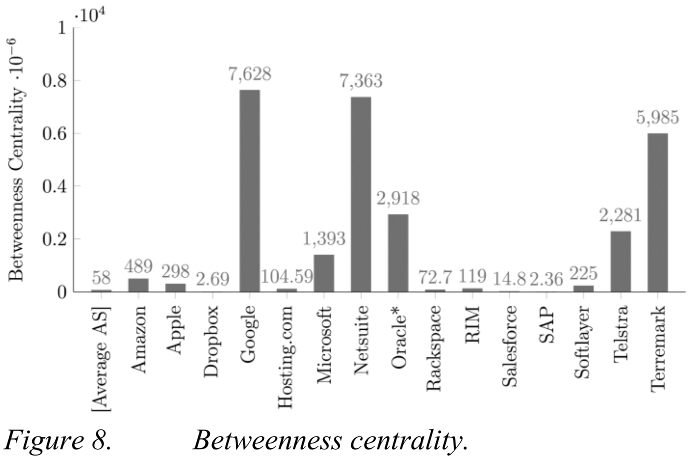

Figure 8. *Betweenness centrality.*

5.2.2.4. Closeness Centrality

The higher the CLC value, the closer a vertex is to the center of the network. Standardized CLC values for each primary AS are shown in Fig. 9. The primary ASs of Terremark, Google, and Oracle are closest to the topological graph center. In total, 51 out of the 78 investigated CSP ASs are closer to the center in terms of CLC than an average vertex in the entire AS dataset.

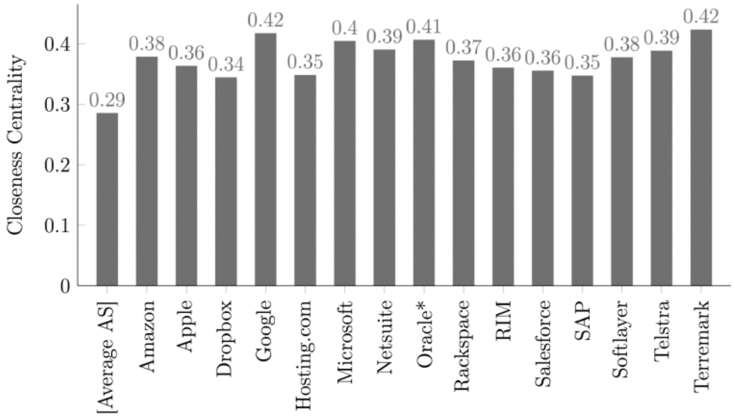

Figure 9. Closeness centrality.

5.2.2.5. Local Average Neighbor Degree (LAND)

LAND summarizes the average degree of all vertices in the one-hop neighborhood of a particular vertex. It makes a statement about the connectedness of the neighborhood of a vertex to the overall graph. This is important because a better-connected vertex has more alternative paths to choose from when one path fails. A node of the full AS-level graph has a LAND of 703.29 on average and would be considered well connected in that respect because every neighbor on average has 703 paths to choose from. However, some caveats apply with respect to possible distortions from the scale-free degree distribution. LAND scores for all primary ASs are shown in Fig. 10. Only six primary ASs (32 of all CSP ASs) have a LAND greater than an average vertex. Though the values do not seem critically low, specific routes and shortest paths could be adversely affected by outages not directly at the CSP but at edges of a low-degree neighbor.

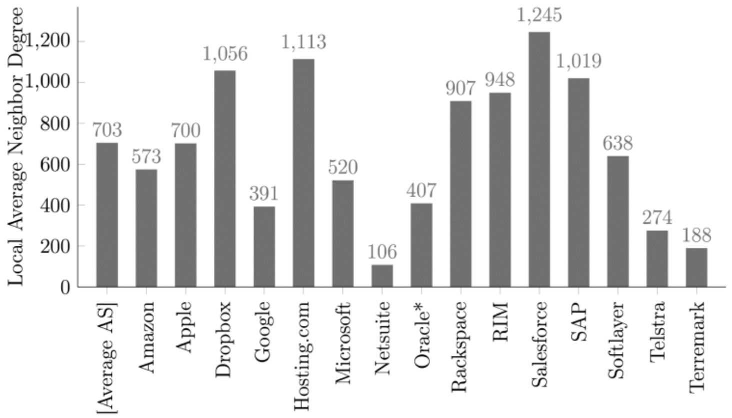

Figure 10. Local average neighbor degree.

5.2.2.6. Local Node Connectivity

In order to utilize LNC, it is required to specify a source vertex and a target vertex. For choosing source vertices, three locations around the globe were selected that will also serve as starting points for traceroute analyses. The first location is Germany (AS 31334, ISP Kabel Deutschland). The second location is Dallas, U.S., where the website network-tools.com is tracing from AS 17819. The third traceroute starting point is Australia, using AS 1221 to trace from the ISP Telstra (telstra.com). These locations were chosen for investigating how geographic distance between vertices influences connectivity. They were kept fixed in order to make the measurements reproducible and comparable. For all source ASs, it was verified whether they are nonadjacent to all target ASs (with the exception of Telstra to Telstra measurements) in order to avoid trivial paths.

LNC calculates the minimum number of vertices that must be removed in order to completely disconnect source and target vertices (NetworkX, 2014). Results are displayed in Table 7. When starting from source vertex AS 31334 (Germany), the amount of vertices that have to be removed in order to disconnect it from some of the CSPs is clearly the target AS degree, e.g., for Dropbox 17, Salesforce 19, Apple 33, and SAP 38. But for Microsoft, Google, Oracle, Softlayer, Terremark, and Netsuite, the LNC is 95 which is much lower than the target AS degree. For source AS 17819 (U.S.), the results are similar. But only 63 vertices need to be removed to disconnect Microsoft, Google, Amazon, Rackspace, Softlayer, Terremark, Netsuite, and Oracle. From source AS 1221 (Australia), in most of these cases 85 nodes would have to be removed, except for Oracle and Rackspace where 88 and 67 are needed, respectively. The minimum number necessary for disconnecting a primary AS is only 11 for Hosting.com, which is even lower than its degree of 14.

It is remarkable that seven CSPs are disconnected by removing the same amount of vertices from each destination. In order to find possible interdependencies between those CSPs of our study, their ASs were removed successively in a simulation according to their degree ranking. Terremark was "removed" first, followed by Netsuite, Google, Oracle, Microsoft, Amazon, and Softlayer. When each single CSP and all of its connections had been removed, the same graph analysis was repeated. We found that the lower the degree of the removed vertex, the smaller the influence of its removal. Even the removal of Terremark, despite its high degree of 897, had the effect that the degrees of Google, Oracle, Microsoft, Telstra, Amazon, Softlayer, Rackspace, RIM, and SAP decreased by only one. The shortest path length for Google went up by 0.16. Centrality measures were left unchanged or the change was very small. The lower the degree of the removed vertex, the less CSPs were affected and also to a smaller extent. This analysis supports the hypothesis that the majority of the investigated CSPs are very well interconnected and there are several paths to choose from, in the case that one CSP is unreachable. This could be regarded as an indicator that multi-cloud approaches (Fabian et al., 2015) across several CSPs could indeed improve reliability not only with respect to data redundancy but also network connectivity and reliability.

5.3. Traceroute Experiments

We conducted traceroute experiments in order to validate our findings by studying CSP connectedness, also at the Internet Protocol layer. For this purpose, the IP addresses used for CS were identified with the help of several WHOIS services. In cases where the CS uses its own dedicated host name, e.g., windowsazure.microsoft.com, the address was used as trace target. Table 5 presents the providers that were traced, along with their IP addresses or host names; here, AS Path shows the ASs traversed and IP Hops denotes the length of the IP routing path.

The traceroute measurements were initially conducted from three different locations on the globe, in order to see how the connectivity differs from varying regions (Germany, United States, Australia). The results serve as an additional measure that quantifies another aspect of CSP connectedness, i.e., the length of IP paths necessary for accessing the service. Each trace was performed repeatedly on different days. The corresponding variation between results was minimal and therefore negligible. Traces for Amazon and Microsoft were aborted from every source because of missing answers from the target routers, which may have been configured not to reply. With Germany and the United States as starting points, successful traces that reached the desired ASs were collected for Oracle, Google, Salesforce, Hosting.com, Rackspace, Softlayer, Terremark, Dropbox, and Netsuite. For Australia, only two traces were recorded completely and reached the desired AS (Google and Netsuite). Other traces from this location either timed out or did not reach the desired AS, which makes them not comparable. Therefore, Australia was excluded from the current analysis.

Since our AS-level graph provides only "high-level" hops between ASs, the IP addresses have been mapped to their associated AS before the analysis was conducted. This mapping also facilitates a comparison with the SSSPL computed from the AS graph. In all cases, the amount of IP hops in the traces from Germany exceeds those starting from the United States, as was to be expected. The amount of AS hops in the cases of Oracle and Google were smaller (1 hop) than from the US (2 hops). For all other traced paths, the amount of AS hops from Germany equals that from the US. A comparison with the SSSPL shows that from these two selected sources, each trace took a slightly shorter path than the average length of the shortest paths from the CSP to all other vertices in the graph. However, since both source ASs for traceroute belong to ISPs that are better connected than an average AS, we believe that these experiments present selective evidence that the AS-based graph analysis does indeed deliver qualitatively valid and meaningful results.

CSP	Trace target	ASNs (GER)	ASNs (US)	Hops (GER)	Hops (US)	SSSPL
Oracle*	cloud.oracle.com	31,334–20,940	36,351–19,108–20,940	9	4	2.46
Google	apps.google.com	31,344–15,169	36,351–3356–15,169	15	6	2.40
	appengine.google.com	31,344–15,169	36,351–3356–15,169	15	6	2.40
Salesforce	204.14.234.33	31,344–3356–14,340	36,351–3356–14,340	23	11	2.82
Hosting.com	hosting.com	31,344–3356–20,021	36,351–3356–20,021	21	11	2.87
Rackspace	173.203.44.122	31,344–3356–19,994	36,351–3356–19,994	30	10	2.69
Softlayer	66.228.118.53	31,344–1200–36,351	36,351–3356–36,351	16	6	2.65
Terremark	66.165.161.46	31,344–1200–23,148	36,351–3356–23,148	15	10	2.37
Dropbox	199.47.217.179	31,344–3356–19,679	36,351–3356–19,679	22	10	2.91
Netsuite	167.216.129.12	31,344–3356–3561	36,351–3356–3561	21	8	2.56

Table 5. Results of the traceroute experiments.

6. Discussion

6.1. Interpretation of Results

Table 6 presents a synopsis of results for the primary CSP ASs. Table 8 (Table A in appendix) also shows the results for the other CSP ASs of our sample, which in general display much weaker connectivity scores as has been discussed in the preceding section. Furthermore, average scores for the entire AS graph are displayed for comparison.

All of the analyzed CSP ASs, both primary and non-primary, are slightly closer to the topological center than an average AS according to both ECC and SSSPL. Also with respect to the DC, the primary ASs were found to be significantly better connected than an average vertex in the graph. According to DC, Terremark's primary AS is the best connected vertex. EC for all primary ASs is above average, but for some CSPs only slightly so. Terremark has the highest EC score followed by Oracle and Google. Less than half of all analyzed vertices have an EC greater than an average vertex, indicating again a very different connectivity of non-primary AS.

For every CSP, the primary AS has the highest BC. The most important vertices in terms of BC are AS 15169 of Google and AS 3561 of Netsuite (0.0074). For 34 of all investigated vertices, the BC score is zero, which means that no shortest path is leading through them, indicating that those ASs are not important as transit AS for the communication between others. This could be relevant for cloud-based data sharing between globally distributed clients. In total, all primary ASs (and 88% of all investigated ASs) are closer to the center in terms of CLC than an average vertex in the entire AS dataset. The primary ASs of Terremark, Google, and Oracle are closest to the topological graph center. LAND indicates whether a vertex prefers to connect to vertices with high or low degrees (Mahadevan et al., 2006). Results for the primary ASs are also shown in Table 6. Only six out of those ASs have a LAND greater than an average vertex. Therefore, specific routes and shortest paths could be adversely affected by outages at low-degree neighbor.

CSP	ASN	ECC	SSSPL	Degree	EC	BC (·10-6)	CLC	LAND
Amazon	16,509	7	2.6486	145	0.02767	488.83	0.3775	573.4
Apple	714	7	2.7546	33	0.00332	298.48	0.3630	700.2
Dropbox	19,679	7	2.9083	17	0.00424	2.69	0.3438	1055.5
Google	15,169	6	2.3994	326	0.03971	7627.53	0.4168	391.0
Hosting.com	20,021	7	2.8739	14	0.00183	104.59	0.3480	1112.5
Microsoft	8075	7	2.4726	204	0.03465	1392.80	0.4044	519.9
Netsuite	3561	7	2.5644	486	0.01091	7362.74	0.3900	106.4
Oracle*	20,940	6	2.4617	285	0.04000	2918.41	0.4062	407.1
Rackspace	15,395	7	2.6877	67	0.01937	72.72	0.3721	906.9
RIM	18,705	7	2.7760	52	0.01700	119.06	0.3602	947.7
Salesforce	14,340	7	2.8184	19	0.00507	14.79	0.3548	1244.9
SAP	12,510	7	2.8839	38	0.01422	2.36	0.3467	1019.2
Softlayer	36,351	7	2.6547	131	0.03055	224.70	0.3767	638.1
Telstra	4637	6	2.5794	217	0.01692	2280.71	0.3877	273.59
Terremark	23,148	7	2.3666	897	0.09928	5984.73	0.4225	187.9
AS graph	(All)	7.83	3.56	8.97	0.00096	58.00	0.2851	703.3

Table 6. Results for distance and centrality measures.

The results of centrality and distance measures also indicate that Terremark, Google, Netsuite, Oracle, Microsoft, Amazon, and Softlayer are close to the center and well interconnected. Therefore, these CSPs are robust in case of network failures due to the huge amount of paths they can choose from. CSPs located rather at the topological border (Rackspace, RIM, SAP, Salesforce, Apple, Dropbox, Hosting.com) could potentially be more prone to failure. In general, the outage of one cloud provider does not significantly affect the overall connectivity of other CSPs; therefore, judging from the topological structure, cascading failures are not to be expected.

CSP	ASN	Degree	LNC:GER	US	AUS
Amazon	16,509	145	95	63	85
Apple	714	33	33	33	33
Dropbox	19,679	17	17	17	17
Google	15,169	326	95	63	85
Hosting.com	20,021	14	11	11	11
Microsoft	8075	204	95	63	85
Netsuite	3561	486	95	63	85
Oracle*	20,940	285	95	63	85
Rackspace	15,395	67	67	63	67
RIM	18,705	52	51	51	51
Salesforce	14,340	19	19	19	19
SAP	12,510	38	38	38	38
Softlayer	36,351	131	95	63	85
Telstra	1221	200	85	66	200
Terremark	23,148	897	95	63	85

Table 7. Results for global connectivity (LNC).

For investigating end-to-end connectivity of cloud ASs around the globe, we found that LNC is a relevant measure and provides another perspective. With low-degree CSPs, the amount of vertices that have to be removed in order to disconnect them from our measurement sources is clearly equal to the target AS degree (Dropbox, Salesforce, Apple, SAP) or lower for (Hosting.com). On the other hand, the LNC for Microsoft, Google, Oracle, Softlayer, Terremark, and Netsuite is higher, although still much lower than the target AS degree. This again underlines the importance of considering the deeper Internet structure for connectivity beyond the degree.

In summary, the better connected a CSP and the closer it is located to the center of the AS topology, the higher its topological resilience. Furthermore, it is to some extent possible to compensate for a low degree and a larger distance to the center with a high average neighbor degree. A larger distance from the center may not necessarily affect the resilience if the CSP is connected to many important neighbors. According to the ranking resulting from our graph analysis and subject to possible limitations of the data set, Terremark and Google could be considered as some of the better-connected CSPs in the sample. Other CSPs are clearly seen to rank at the bottom in this comparison with respect to their topological connectedness, but they are still better connected than an average AS. For nearly every measure and CSP, non-primary ASs are not as well connected and would generally not be able to replace primary ASs without substantial reconnection efforts.

6.2. Limitations and Future Work

In order to keep a narrow focus, our article deliberately does not discuss the technical or economic perspective of potential adversaries. We also do not discuss corresponding countermeasures in the current article. Our analysis is based on a rich AS-level dataset integrated from several sources but still constitutes an incomplete snapshot of the Internet. This limitation can be addressed by repeated and possibly automated measurements for recent AS and IP-level datasets. Moreover, we aim to study the effects of economic constraints such as those caused by policy-driven routing (Dolev et al., 2006) in the future. Though all metrics we chose are relevant for connectivity, they display complex interactions and sometimes partial correlations on different graphs; their exact and objective weighting is an open question for future work, as are simulative studies on heterogeneous graph models.

6.3. Management Recommendations

For CSPs that so far have achieved a lower position in our connectivity ranking, their first step would be to reassess the situation based on private knowledge of internal connectivity or backup links that may not be visible to the public. Then, a general recommendation for CSPs to gain better connectivity results would be to increase the amount of direct- and next-hop connections. This could be achieved in various ways. They could contract and connect to other ASs (increasing DC), preferring those with high degree (i.e., increasing LAND) and excellent centrality in terms of EC, BC, and CLC. Furthermore, the CSP could also become a customer of different providers offering better options with respect to connectivity and availability. Another promising strategy would be to directly connect to additional major Internet Exchange Points in order to connect with multiple ASs at the same time.

For cloud customers, a good strategy for access to a reliable service could be to benchmark and select CSPs with respect to their connectivity characteristics, complementing frameworks for CS rankings such as those put forth by Garg et al. (2013). One guideline for supporting CSP selection is depicted in Fig. 11. Since high availability may be only one important goal among others, several CSP candidates can be selected according to further business requirements. For these CSPs, the ASs need to be identified and current graph data on Internet connectivity must be gathered and integrated as presented in Section 4.1. These and the following steps could also be part of an independent service and could be integrated into certification processes for CSPs (Lansing et al., 2013). During connectivity analysis, DC should be used as the first filtering requirement because it reflects an important measure of direct connectivity to neighboring ASs. This could reduce the set of CSP candidates. In the next step, the customer should estimate from what locations his or her service would be used in the future. Where are the most important clients located, both in terms of real-word geography and Internet graph topology? This can be based on geography-to-IP and IP-to-AS mappings, and can also be provided by the connectivity analysis service.

Based on the estimated fractions of "near" (in the AS topology) and global business, subjective weights for the following metrics can then be derived. For global business, high BC and CLC values would be important. For local clients, LAND and EC are important metrics. Note that this emphasis of certain metrics can be modified by experience. If particular countries or regions of clients of the CS are known, several LNC analyses based on corresponding source ASs should also be conducted. The results of all analyses are weighted according to the expected fraction of business and aggregated. Using these scores, a few final CSP candidates can be selected. For each of them, further in-depth IP-level analyses can be conducted to verify their current connectivity. In this step, CSPs should also be consulted directly in order to get more information about potential backup links that have so far not been discovered. Based on this information, the final CSP can be selected.

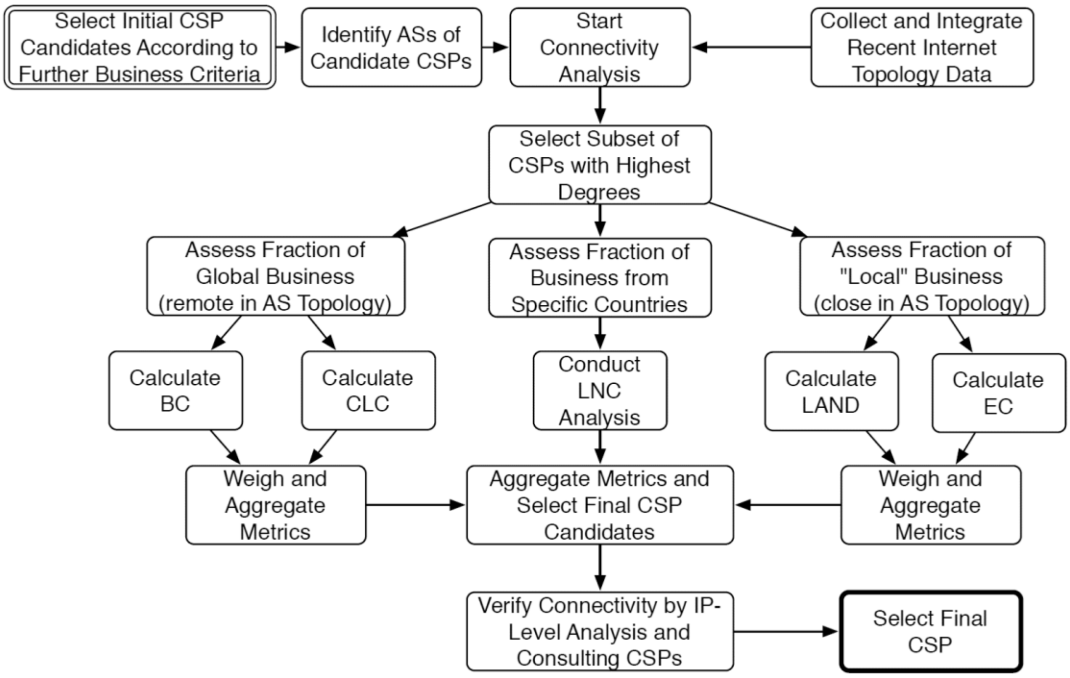

Figure 11. Support for CSP selection.

Reliability in general is a complex and multidimensional issue that depends not only on topological aspects. Another promising strategy for CSP customers would be to not rely just on a single well-connected service but to distribute the CS using several CSPs and ASs for risk diversification using a multi-cloud approach. Furthermore, formalized approaches to modeling and assessing cloud availability, such as those proposed by Frattini et al. (2014), could be extended to address the topological connectivity issues discussed in this article.

7. Conclusion

This article investigates the topological connectivity of top-ranked CSPs. Based on modeling of the Internet topology at the AS level, metrics from graph theory provide information about the structural vulnerability of CSPs to Internet outages. We discovered that CSPs differ by a large extent in the connectivity measures we applied, sometimes by more than an order of magnitude. Our methods can be used by CSPs to assess and possibly improve their network connectivity, and can also support a customer's process of CS benchmarking and selection.

Appendix

CSP (remark)	ASN	ECC	SSSPL	Degree	EC	BC (10^{-3})	CLC	LAND
Amazon (EU)	8987	8	2.63	1	0.00012	0	0.2741	145
Amazon (AWS)	14,618	8	3.64	2	0.00012	0.000039	0.2746	84.5
Amazon (AWS)	16,509	7	2.65	145	0.02767	0.488827	0.3775	573.4
Amazon	38,895	7	3.15	10	0.00118	0.000231	0.3173	551.9
Amazon (EU)	39,111	7	3.32	8	0.00141	0.000729	0.3311	948.6
Apple (iCloud)	714	7	2.75	33	0.00332	0.298477	0.363	700.2
Apple (Store)	2709	7	3.04	5	0.00052	0.000476	0.3288	1450
Dropbox	19,679	7	2.91	17	0.00424	0.002686	0.3438	1055.5
Dropbox	54,372	7	3.09	2	0.00055	0	0.3236	2685.5
Google (Gmail)	15,169	6	2.4	326	0.03971	7.627524	0.4168	391
Google (Fiber)	16,591	7	3.09	2	0.00046	0	0.3239	1068.5
Google (Fiber)	19,448	8	3.21	2	0.00045	0	0.3114	2709.5
Google (AdMob)	22,577	7	2.96	12	0.00298	0.00066	0.3373	1042.1
Google	24,424	7	3.39	4	0.00019	0.036732	0.2946	119.5
Google (YouTube)	36,040	7	2.73	30	0.00683	0.036384	0.3662	1084
Google (WiFi)	36,492	7	2.95	3	0.0012	0	0.3392	3178.7
Google (EU)	41,264	7	3.01	4	0.00095	0.000056	0.3323	1742.5
Google (APAC)	45,566	7	3.4	1	0.00018	0	0.2942	326
Hosting.com	4378	8	3.36	2	0.00015	0	0.2979	1759.5
Hosting.com	20,021	7	2.87	14	0.00183	0.104592	0.348	1112.5
Microsoft	3598	7	3.1	2	0.00052	0	0.3223	2218.5
Microsoft	5761	8	3.47	1	0.00015	0	0.288	204
Microsoft	6182	7	3.85	1	0.00002	0	0.2596	101
Microsoft (EU)	8068	7	2.98	39	0.00531	0.006501	0.3358	376.3
Microsoft	8069	7	2.91	31	0.00381	0.007469	0.3438	564.2
Microsoft	8072	7	3.96	2	0.00001	0	0.2524	57
Microsoft	8075	7	2.47	204	0.03465	1.3928	0.4044	519.9
Microsoft	13,811	8	3.47	1	0.00006	0	0.2884	2333
Microsoft	20,046	7	3.3	4	0.00027	0.00176	0.3032	882.8
Microsoft (XBox)	23,468	7	3.1	2	0.00052	0	0.3223	2218.5
Microsoft	36,006	7	3.11	1	0.00037	0	0.3212	4233
Netsuite	3561	7	2.56	486	0.01091	7.362742	0.39	106.4
Netsuite	14,919	8	3.21	3	0.00033	0	0.312	1326
Oracle	792	7	3.06	2	0.00044	0	0.3266	2972
Oracle	793	7	3.11	2	0.00038	0.000593	0.3213	2132.5
Oracle	794	7	2.97	5	0.00073	0.000166	0.3362	1884
Oracle	1215	7	3.1	3	0.00047	0	0.3228	1520.3
Oracle	1217	8	4.08	1	0.00001	0	0.245	95
Oracle* (Akamai)	20,940	6	2.46	285	0.04	2.918409	0.4062	407.1
Oracle (JAP)	23,885	8	3.54	2	0.00012	0	0.2824	275
Oracle (INDIA)	38,358	7	3.93	2	0.00001	0	0.2542	93.5
Oracle	41,900	8	3.81	9	0.00002	0.00125	0.2626	22.8
Oracle (EMEA)	52,019	7	3.1	4	0.00053	0	0.3031	597.5
Rackspace	10,532	7	3.03	5	0.00056	0.00028	0.3301	1567.2
Rackspace	12,200	7	2.99	2	0.00071	0	0.3346	4281.5
Rackspace (UK)	15,395	7	2.69	67	0.01937	0.072718	0.3721	906.9
Rackspace	19,994	7	2.86	14	0.00196	0.003352	0.3494	1103
Rackspace (AUS)	22,720	8	3.45	2	0.0001	0	0.2901	934
Rackspace	27,357	7	2.91	10	0.00159	0.001606	0.3438	1306.4
Rackspace	33,070	7	2.87	17	0.00214	0.002063	0.3479	891.2
Rackspace	36,248	7	3	24	0.00468	0.002109	0.3335	600.8
Rackspace	45,187	7	3.2	6	0.00101	0	0.313	624
RIM	9798	8	3.36	3	0.00018	0.000315	0.2976	306
RIM	13,340	7	3.07	4	0.00053	0.000263	0.3255	1268.3
RIM	18,705	7	2.78	52	0.017	0.119063	0.3602	947.7
RIM	26,281	7	3.02	5	0.0006	0.00092	0.3312	1481
RIM	31,974	7	3.07	2	0.00052	0	0.3255	2524
RIM	41,433	7	3.09	2	0.00047	0	0.3233	2482
RIM	47,454	8	3.57	3	0.00025	0.00135	0.2802	193.3
RIM (BELG)	51,867	7	3.07	2	0.00052	0	0.3255	2524

RIM (CAN)	55,721	8	3.78	1	0.00008	0	0.2648	52
Salesforce	14,340	7	2.82	19	0.00507	0.014794	0.3548	1244.9
Salesforce	45,422	7	3.22	4	0.00053	0	0.3109	650
SAP (US)	6979	7	2.96	7	0.00076	0.026431	0.3376	1376
SAP (GER)	12,510	7	2.88	38	0.01422	0.002361	0.3467	1019.2
SAP (ASIA)	18,219	8	3.64	2	0.00006	0	0.2748	133
Softlayer	36,351	7	2.65	131	0.03055	0.224692	0.3767	638.1
Telstra	1,221	7	2.91	200	0.00355	4.344254	0.3439	81
Telstra (EU)	1290	7	2.68	154	0.02799	0.370339	0.3725	461.2
Telstra	4,637	6	2.58	217	0.01692	2.280711	0.3877	273.59
Telstra	57,486	8	3.72	2	0.00013	0	0.2687	99
Terremark	11,303	8	3.37	1	0.00044	0	0.297	897
Terremark	21,533	8	3.37	1	0.00044	0	0.297	897
Terremark	23,148	7	2.37	897	0.09928	5.984726	0.4225	187.9
Terremark (CAR)	27,991	8	3.36	3	0.00044	0.000673	0.298	303.67
Terremark (COL)	27,992	8	3.12	5	0.0011	0.004337	0.3219	890.8
Terremark	28,625	8	3.18	47	0.0011	0.135079	0.3143	162.4
Terremark	50,146	8	3.31	7	0.00071	0.001898	0.3017	247.3

Table A. Metrics for all CSP ASs (sample size: 78).

References

Abu-Libdeh, H., Princehouse, L., and Weatherspoon, H. (2010). RACS: A case for cloud storage diversity. In Proceedings of the 1st ACM symposium on cloud computing (SoCC '10) (pp. 229–240). New York, NY, USA: ACM.

Agapi, A., Birman, K., Broberg, R., Cotton, C., Kielmann, T., Millnert, M., et al. (2011). Routers for the cloud: Can the Internet achieve 5-nines availability? IEEE Internet Computing, 15(5), 72–77.

Akamai (2001). Oracle and Akamai deliver technology to accelerate web applications and content. http://www.akamai.com/html/about/press/releases/2001/press_043001.html

Albert, R., Jeong, H., and Barabasi, A.-L. (2000). Error and attack tolerance of complex networks. Nature (406), 378–382.

Amazon (June 2013). Amazon S3 service level agreement. http://aws.amazon.com/s3/sla/.

Amazon (June 2013). Amazon EC2 service level agreement. http://aws.amazon.com/ec2/sla/.

ARIN Board of Trustees (2013). American registry for Internet numbers. https://arin.net/.

Barabasi, A., and Albert, R. (1999). Emergence of scaling in random networks. Science, 286(5439), 509–512.

Baumann, A., and Fabian, B. (2014). Who runs the Internet? Classifying autonomous systems into industries. In Proceedings of the 10th international conference on web information systems and technologies (WEBIST), Barcelona, Spain, 2014.

Baumann, A., and Fabian, B. (2014). How robust is the Internet? – Insights from graph analysis. In Proceeding of the 9th international conference on risks and security of internet and systems (CRiSIS 2014). LNCS. Trento, Italy: Springer.

BGP Reports (2013). BGP routing table analysis reports. http://bgp.potaroo.net/.

Bilal, K., Malik, S. U. R., Khan, S. U., and Zomaya, A. Y. (2014). Trends and challenges in cloud data centers. IEEE Cloud Computing Magazine, 1(1), 10–20.

Bilal, K., Manzano, M., Khan, S. U., Calle, E., Li, K., and Zomaya, A. Y. (2013). On the characterization of the structural robustness of data center networks. IEEE Transactions on Cloud Computing, 1(1), 1.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D.-U. (2006). Complex networks: Structure and dynamics. Physics Reports, 424(4–5), 175–308.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. Journal of Mathematical Sociology, 2(1), 113–120.

Bonacich, P., and Lloyd, P. (2001). Eigenvector-like measures of centrality for asymmetric relations. Social Networks, 191–201.

CAIDA (2013). Ark: Archipelago measurement infrastructure. http://www.caida.org/projects/ark/.

Catteddu, D., and Hogben, G. (2009). Cloud computing – Benefits, risks and recommendations for information security. Tech. rep., European Network and Information Security Agency (ENISA).

Cunha, I., Teixeira, R., Veitch, D., and Diot, C. (2011). Predicting and tracking Internet path changes. In Proceedings of the ACM SIGCOMM 2011 conference, 2011 (pp. 122–133).

Data Center Knowledge (2013). Industry news and analysis about datacentres. http://www.datacenterknowledge.com/.

Deng, W., Karaliopoulos, M., Mühlbauer, W., Zhu, P., Lu, X., and Plattner, B. (2011). kFault tolerance of the Internet AS graph. Computer Networks, 55(10), 2492–2503.

Dimitropoulos, X., Krioukov, D., Fomenkov, M., Huffaker, B., Hyun, Y., Claffy, K., et al. (2007). AS relationships: Inference and validation. ACM SIGCOMM Computer Communication Review (CCR), 37(1), 29–40.

Dolev, D., Jamin, S., Mokryn, O., and Shavitt, Y. (2006). Internet resiliency to attacks and failures under BGP policy routing. Computer Networks, 50(16), 3183–3196.

Emerson Network Power (2011). Understanding the cost of data center downtime. http://www.emersonnetworkpower.com/.

Erdös, P., and Renyi, A. (1959). On random graphs I. Publicationes Mathematicae (Debrecen), 290–297

Ermakova, T., Fabian, B., and Zarnekow, R. (2013). Security and privacy system requirements for adopting cloud computing in healthcare data sharing scenarios. In: Proceedings of the nineteenth Americas conference on information systems (AMCIS), Chicago, Illinois, 2013.

Ermakova, T., and Fabian, B. (2013). Secret sharing for health data in multi-provider clouds. In Proceedings of the 15th IEEE conference on business informatics (CBI 2013) (pp. 93–100). IEEE.

Fabian, B., Ermakova, T., and Junghanns, P. (2015). Collaborative and secure sharing of healthcare data in multi-clouds. Information Systems, 48, 132–150.

Frattini, F., Trivedi, K. S., Longo, F., Russo, S., and Ghosh, R. (2014). Scalable analytics for IaaS cloud availability. IEEE Transactions on Cloud Computing, 2(1), 57–70.

Freeman, L. (1977). A set of measures of centrality based on betweenness. Sociometry, 40, 35–41.

Garg, S. K., Versteeg, S., and Buyya, R. (2013). A framework for ranking of cloud computing services. Future Generation Computer Systems, 29(4), 1012–1023.
Gill, P., Jain, N., and Nagappan, N. (2011). Understanding network failures in data centers: Measurement, analysis, and implications. ACM SIGCOMM Computer Communication Review, 41(4), 350–361.
Google (January 2015). Google cloud storage, Google prediction API, and Google BigQuery SLA. https://cloud.google.com/storage/sla.
Guo, C., Wu, H., Tan, K., Shi, L., Zhang, Y., and Lu, S. (2008). DCell: A scalable and faulttolerant network structure for data centers. In Proceedings of the ACM SIGCOMM 2008 conference (pp. 75–86). New York, NY, USA: ACM.
Hagberg, A. A., Schult, D. A., Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In Proceedings of the 7th python in science conference (SciPy2008), Pasadena, CA, USA (pp. 11–15).
Hajjat, M., Sun, X., Sung, Y.-W. E., Maltz, D., Rao, S., Sripanidkulchai, K., et al. (2010). Cloudward bound: Planning for beneficial migration of enterprise applications to the cloud. ACM SIGCOMM Computer Communication Review (CCR), 40(4), 243–254.
Hu, Z., Zhu, L., Ardi, C., Katz-Bassett, E., Madhyastha, H. V., Heidemann, J., et al. (2014). The need for end-to-end evaluation of cloud availability. In M. Faloutsos and A. Kuzmanovic (Eds.), Passive and active measurement. Proceedings of the 15th international conference (PAM 2014), Los Angeles, CA, USA. LNCS (Vol. 8362, pp. 119–130). Springer.
Internet Business Solutions AG (2013). AS number firefox addon. http://www.asnumber.networx.ch/.
IWGCR (2013). International working group on cloud computing resiliency. http://iwgcr.org/.
Kammer, F., and Taubig, H. (2005). Connectivity. In U. Brandes and T. Erlebach (Eds.), Network analysis: Methodological foundations. LNCS (Vol. 3418, pp. 143–177). Springer.
Katz-Bassett, E., Madhyastha, H. V., John, J. P., Krishnamurthy, A., Wetherall, D., and Anderson, T. (2008). Studying black holes in the Internet with hubble. In Proceedings of the 5th USENIX symposion on networked systems design and implementation (NSDI ’08), 2008.
Kong, X. T., Fang, J., Luo, H., and Huang, G. Q. (2014). Cloud-enabled real-time platform for adaptive planning and control in auction logistics center. Computers & Industrial Engineering (0).
Kryvinska, N., Strauss, C., and Zinterhof, P. (2010). A model to provide higher services availability for the mission critical applications. In The international conference on ambient systems, networks and technologies (ANT-2010), in conjunction with iiWAS2010, 8-10 November, Paris, France, 2010 (pp. 221–229).
Kryvinska, N., Strauss, C., and Zinterhof, P. (2011). ‘Variated availability’ approach to the services manageable delivering. In The fifth international conference on innovative mobile and internet services in ubiquitous computing (IMIS-2011), June 30th–July 2nd, 2011, Korean Bible University (KBU), Seoul, Korea, 2011 (pp. 106– 113).
Kryvinska, N., and Strauss, C. (2013). Conceptual model of business services availability vs. interoperability on collaborative IoT-enabled eBusiness platforms. In N. Bessis, F. Xhafa, D. Varvarigou, R. Hill, and M. Li (Eds.), Internet of things and inter-cooperative computational technologies for collective intelligence. Studies in computational intelligence (SCI-460) (pp. 167–187). Berlin Heidelberg: Springer-Verlag.
Landherr, A., Friedl, B., and Heidemann, J. (2010). A critical review of centrality measures in social networks. Business & Information Systems Engineering, 2(6), 371–385.
Lansing, J., Schneider, S., and Sunyaev, A. (2013). Cloud service certifications: Measuring consumer’s preferences for assurances. In Proceedings of the 21st European conference on information systems, Utrecht, Netherlands, 2013.
Li, A., Yang, X., Kandula, S., and Zhang, M. (2010). CloudCmp: Comparing public cloud providers. In Proceedings of the 10th ACM SIGCOMM conference on Internet measurement (IMC ’10). Association for Computing Machinery, Inc..
Mahadevan, P., Krioukov, D., Fomenkov, M., Huffaker, B., Dimitropoulos, X., Claffy, K., et al. (2006). The Internet AS-level topology: Three data sources and one definitive metric. ACM SIGCOMM Computer Communication Review (CCR), 36(1), 17–26.
Manzano, M., Bilal, K., Calle, E., and Khan, S. (2013). On the connectivity of data center networks. IEEE Communications Letters, 17(11), 2172–2175.
Manzano, M., Calle, E., and Harle, D. (2011). Quantitative and qualitative network robustness analysis under different multiple failure scenarios. In Proceedings of the 3rd international congress on ultra modern telecommunications and control systems and workshops (ICUMT 2011) (pp. 1–7). IEEE.
Manzano, M., Marzo, J., Calle, E., and Manolovay, A. (2012). Robustness analysis of real network topologies under multiple failure scenarios. In Proceedings of the 17th European conference on networks and optical communications (NOC 2012) (pp. 1–6). IEEE.

Markopoulou, A., Iannaccone, G., Bhattacharyya, S., Chuah, C.-N., Ganjali, Y., Diot, C., et al. (2008). Characterization of failures in an operational IP backbone network. IEEE/ACM Transactions on Networking, 16(4), 749–762.
Mell, P., and Grance, T. (2011). The NIST definition of cloud computing. NIST special publication 800-145.
Merit Network Inc. (2013). Internet routing registry. http://www.irr.net/.
Microsoft (December 2014). Service level agreements for most azure services. http://azure.microsoft.com/en-us/support/legal/sla/.
Microsoft Technet (2005). Internet control message protocol (ICMP). http://technet.microsoft.com/.
Motoyama, M., Meeder, B., Levchenko, K., Voelker, G. M., and Savage, S. (2010). Measuring online service availability using Twitter. In Proceedings of the 3rd workshop on online social networks (WOSN'10), USENIX, 2010 (pp. 13–13).
NetworkX (2014). High-productivity software for complex networks. Project documentation. http://networkx.github.io/.
Neumayer, S., and Modiano, E. (2010). Network reliability with geographically correlated failures. In Proceedings of IEEE INFOCOM '10 (pp. 1–9). IEEE.
Neustar Inc. (2013). UltraTools. https://www.ultratools.com/.
Newman, M. (2003). The structure and function of complex networks. SIAM Review, 45(2), 167–256.
Oggerino, C. (2001). High availability network fundamentals: A practical guide to predicting network availability. Cisco Press.
Okamoto, K., Chen, W., and Li, X.-Y. (2008). Ranking of closeness centrality for largescale social networks. In Frontiers in algorithmics. LNCS (Vol. 5059, pp. 186–195). Springer
Panettieri, J. (2013). Top 100 cloud services providers: 2013 edition. Tech. rep., Talkin' cloud. http://talkincloud.com/tc100/.
Quan, L., Heidemann, J., and Pradkin, Y. (2013). Trinocular: Understanding Internet reliability through adaptive probing. In Proceedings of the ACM SIGCOMM 2013 conference (pp. 255–266). ACM.
Renesys (2002). Internet worms and global routing instabilities. http://www.renesys.com/tech/presentations/pdf/renesys-spie2002.pdf.
Renesys (2003). Impact of the 2003 blackouts on Internet communications. http://www.renesys.com/tech/reports/Renesys_BlackoutReport.pdf.
Renesys (2005). Impact of Hurricane Katrina on Internet infrastructure. http://www.renesys.com/tech/presentations/pdf/Renesys-Katrina-Report-9sep2005.pdf.
Renesys (2007). Quaking tables – The Taiwan earthquakes and the Internet routing table. http://www.renesys.com/tech/presentations/pdf/ripe-quaking-tables.pdf.
RIPE (2013). Routing information service. http://www.ripe.net/data-tools/stats/ris/routing-information-service/.
Santhakumaran, A. (2010). Center of a graph with respect to edges. SCIENTIA, Series A: Mathematical Sciences, 19, 13–23.
Schneider, C., Moreira, A., Andrade, J., Jr., Havlin, S., and Herrmann, H. (2011). Mitigation of malicious attacks on networks. Proceedings of the National Academy of Sciences, 108(10), 3838–3841.
Smith, P., Hutchison, D., Sterbenz, J., Schöller, M., Fessi, A., Karaliopoulos, M., et al. (2011). Network resilience: A systematic approach. IEEE Communications Magazine, 49(7), 88–97.
Sterbenz, J. P. G., Hutchison, D., Cetinkaya, E. K., Jabbar, A., Rohrer, J. P., Schöller, M., et al. (2010). Resilience and survivability in communication networks: Strategies, principles, and survey of disciplines. Computer Networks, 54(8), 1245–1265.
Subashini, S., and Kavitha, V. (2011). A survey on security issues in service delivery models of cloud computing. Journal of Network and Computer Applications, 34(1), 1–11.
Toosi, A. N., Calheiros, R. N., and Buyya, R. (2014). Interconnected cloud computing environments: Challenges, taxonomy, and survey. ACM Computing Surveys, 47(1), 7:1–7:47. doi: 10.1145/2593512.
Turing Technology Services (2013). Networktools. http://networktools.nl/.
UCLA (2013). Internet topology collection. http://irl.cs.ucla.edu/topology/.
University of Oregon (2013). Route views project. http://www.routeviews.org/.
van Beijnum, I. (2002). BGP – Building reliable networks with the border gateway protocol. O'Reilly.
Wu, Z., Butkiewicz, M., Perkins, D., Katz-Bassett, E., and Madhyastha, H. V. (2013). SPANStore: Cost-effective geo-replicated storage spanning multiple cloud services. In Proceedings of the twenty-fourth ACM symposium on operating systems principles (SOSP '13) (pp. 292–308). New York, NY, USA: ACM.
Wu, J., Zhang, Y., Morley Mao, Z., and Shin, K. G. (2007). Internet routing resilience to failures: Analysis and implications. In Proceedings of the 2007 ACM CoNEXT conference, New York, NY, USA, 2007.
Wu, D., Greer, M. J., Rosen, D. W., and Schaefer, D. (2013). Cloud manufacturing: Strategic vision and state-of-the-art. Journal of Manufacturing Systems, 32(4), 564–579.
Xiao, S., Xiao, G., and Cheng, T. H. (2008). Tolerance of intentional attacks in complex communication networks. IEEE Communications Magazine, 46(1), 146–152.

2 Management Perspective: E-Commerce

ARTICLE 7:

MAXIMIZE WHAT MATTERS: PREDICTING CUSTOMER CHURN WITH DECISION-CENTRIC ENSEMBLE SELECTION

Authors: Annika Baumann, Humboldt-Universität zu Berlin, Berlin, Germany

Stefan Lessmann, Humboldt-Universität zu Berlin, Berlin, Germany

Kristof Coussement, IÉSEG School of Management, Lille, France

Koen W. De Bock, IÉSEG School of Management, Lille, France

Proceedings: 23rd European Conference on Information Systems (ECIS'15)

Publication year: 2015

Abstract

Churn modeling is important to sustain profitable customer relationships in saturated consumer markets. A churn model predicts the likelihood of customer defection. This is important to target retention offers to the right customers and to use marketing resources efficiently. The prevailing approach toward churn model development, supervised learning, suffers an important limitation: it does not allow the marketing analyst to account for campaign planning objectives and constraints during model building. Our key proposition is that creating a churn model in awareness of actual business requirements increases the performance of the final model for marketing decision support. To demonstrate this, we propose a decision-centric framework to create churn models. We test our modeling framework on eight real-life churn data sets and find that it performs significantly better than state-of-the-art churn models. Further analysis suggests that this improvement comes directly from incorporating business objectives into model building, which confirms the effectiveness of the proposed framework. In particular, we estimate that our approach increases the per customer profits of retention campaigns by $.47 on average.

Keywords: *Predictive Analytics, Churn Modelling, Marketing Decision Support, Ensemble Selection*

1. Introduction

Today, managers are more than ever interested to build enduring customer relationships (e.g., Fader and Hardie, 2010). Acquiring new customers in saturated markets is challenging, and more expensive than retaining existing customers (e.g., Bhattacharya, 1998; Colgate and Danaher, 2000). Moreover, long-term customers generate higher profits, are less sensitive to competitive actions, and may act as promoters through word of mouth (e.g., Ganesh et al., 2000; Reichheld, 1996; Zeithaml et al., 1996). Although relationship management instruments such as loyalty programs often reduce churn (e.g., Kopalle et al., 2012; Lewis, 2004; Verhoef, 2003), customer attrition remains a major threat to the financial health of many companies (e.g., Risselada et al., 2010; Schweidel et al., 2008; Thomas et al., 2004). For example, T-Mobile USA lost half a million of its most lucrative customers in the first quarter of 2012 (Bensinger and Tibken, 2012). Targeted marketing actions using retention campaigns toward risky customers can significantly reduce churn rates and increase firm profits (Burez and Van den Poel, 2007).

To do so, marketing analysts can choose from a variety of approaches to build predictive models that estimate the probability of a customer to become a churner. The choice of the modeling technique is important because it has a direct impact on prediction quality and thus on the profitability of all subsequent targeted marketing efforts (e.g., Neslin et al., 2006; Risselada et al., 2010). Many studies have thus compared different methods to identify a 'best' churn modeling technique (e.g., Verbeke et al., 2012).

Previous churn modeling techniques embody the standard philosophy toward predictive learning: maximize the fit between the model and historical data using statistical quality criterion such as the likelihood. We argue that focusing only on statistical accuracy may be misleading. Marketers use churn models to aid resource allocation decisions. If a marketing budget facilitates soliciting *N* customers with a retention program, the churn model's task is to identify the top-*N* customers with the highest attrition risk. Conventional churn modeling techniques are agnostic of this context. They ignore the budget constraint and tend to assess 'fit' across all customers, rather than emphasizing accuracy among recipients of the campaign. However, research on marketing decision support systems suggests that a mismatch between the actual decision task (resource allocation) and its representation in a decision support model (e.g., likelihood maximization to build a churn model) has a negative impact on decision outcomes and performance (e.g., Lilien, 2011). Therefore, our key proposition is that creating churn models in awareness of business requirements and objectives improves the quality of resource allocation decisions and thus the profitability of retention activities. To test our proposition, we develop a decision-centric churn modeling framework on the basis of a recent machine learning approach called ensemble selection (e.g., Partalas et al., 2010). Using this methodology, we create churn models that explicitly maximize the lift index, which is a well-established measure to assess campaign planning models. We call this approach decision-centric ensemble selection (DCES) because it emphasizes the ultimate decision problem during model building. To explore the effectiveness of DCES in a systematic way, we compare it to alternative churn modeling approaches and analyze the following research questions.

- RQ1: Does DCES outperform the popular logit choice model?
- RQ2: How does DCES perform in relation to advanced single classifiers?
- RQ3: Can DCES beat sophisticated ensemble learners?
- RQ4: Does our lift-based modeling philosophy explain the performance of DCES?

We organize the remainder of the article as follows: In the next section, we provide an overview of the related literature. We then discuss the lift index as a measure of resource allocation efficiency, before we present our DCES framework. Next, we describe the data sets employed in our study and answer our research questions. Afterwards, we conclude the paper with a discussion of findings and implications.

2. Related Literature

Modeling customer churn is part of retention management (e.g., Musalem and Joshi, 2009). In general, we distinguish two groups of churn models, explanatory and predictive models. Approaches of the first category develop models to explain churn patterns on the basis of various constructs, including the firms' marketing activities (Lewis, 2004), customer knowledge (Capraro et al., 2003), or attitudinal concepts such as satisfaction (Bolton, 1998; Gustafsson et al., 2005) or perceived quality (Zeithaml et al., 1996). Approaches that model churn probabilities in a time-dependent manner using [NBD]/Pareto or Markov models (e.g., Gupta and Zeithaml, 2006) also belong to this category. Prediction models follow a data-driven modeling paradigm and are often opaque. Their advantage is that they are explicitly designed for forecasting purposes and typically predict more accurately than explanatory models (Shmueli and Koppius, 2011). Predictive accuracy is important in marketing research and practice (Cui and Curry, 2005). It is especially important in churn modeling to target retention offers to the right customers and to use marketing resources efficiently (e.g., Lemmens and Croux, 2006; Neslin et al., 2006).

Several methods and algorithms are available to develop predictive churn models such as partial least squares (Kim et al., 2013) and those based on supervised learning (e.g., Hastie et al., 2009). For example, the logit choice model is widely used in industry (Cui and Curry, 2005) and has been shown to perform relatively well when compared to more advanced techniques (e.g., Neslin et al., 2006; Risselada et al., 2010). However, large-scale benchmarking studies provide evidence that ensemble models, which combine the forecasts of multiple base models, predict customer churn most accurately (e.g., Lemmens and Croux, 2006; Verbeke et al., 2012). Empirical evidence confirms the efficacy of the ensemble paradigm and suggests that combining the predictions from multiple alternative (base) models is a powerful modeling approach in general (e.g., Bhattacharya et al., 2011; Lessmann and Voß, 2010; Loterman et al., 2012). For this reason, we chose the ensemble principle as basis for our DCES framework. In particular, to develop decision-centric churn models, we propose a base model combination strategy that maximizes the profitability of a customer retention campaign (see chapter 4 for details).

3. Performance Measurement

The lift measure is a performance indicator for targeting models (Ling and Li, 1998). It grounds on a list of customers ordered according to their model-estimated churn probabilities (from highest to lowest risk of attrition). We define the lift measure *Ld* for some decile *d* of the ordered list as:

$$L_d = \frac{\hat{\pi}_d}{\hat{\pi}}, \tag{1}$$

where $\hat{\pi}$ and $\hat{\pi}_d$ denote the fraction of actual churners among all customers and those ranked in the top-*d* decile, respectively. Note that a campaign that targets *d* percent of the customers at random will, on average, reach $\hat{\pi}$ actual churners. Therefore, the lift quantifies how much a model improves over a random targeting. In addition, there is a direct link between the lift of a churn model and the profitability of a retention campaign (Neslin et al., 2006). To see this, note that selecting a value for *d* is equivalent to imposing a budget constraint in that it implies a specific campaign size. Furthermore, only actual churners that receive and accept the retention program creates value. This indicates that $\hat{\pi}_d$ is the key driver of campaign profitability. Interested readers are referred to Verbraken et al.

(2012) for a comprehensive discussion of the economics of churn prediction and the lift measure, respectively.

4. Decision-Centric Ensemble Selection

4.1. Motivation and Overview

The prevailing approach to develop a churn model is to use some general-purpose prediction method. Such methods build a model by minimizing some statistical loss function over training data. For example, the logit choice model minimizes the negative log-likelihood, whereas decision tree–based methods use information-theoretic criteria. The analyst can select the prediction method but has little choice in the loss function. Consequently, there is some mismatch between the analyst's objective and the objective function within the prediction method. To better align these two objectives, our DCES framework accounts for business objectives during model building. DCES grounds on a modeling paradigm called ensemble selection. Ensemble selection consists of three stages: (1) constructing a library of candidate models (*model library*), (2) selecting an "appropriate" subset of models for the ensemble (*candidate selection*), and (3) combining the predictions of the chosen models to produce the final (ensemble) forecast (*forecast combination*). Several alternative approaches follow these guidelines and differ mainly in how to organize candidate selection in stage two (e.g., Partalas et al., 2010). The directed hill-climbing strategy (Caruana et al., 2004) is particularly well suited for our purpose because it can accommodate arbitrary accuracy indicators. The following subsections explain the stages of this approach in more detail, and our specific design decisions to develop a churn modeling framework that is driven by actual business objectives.

4.2. Model Library

At first, we construct a large library of candidate churn models. The success of any ensemble strategy depends on the diversity of ensemble members (e.g., Kuncheva, 2004). Our approach to control the error-correlation among candidate models' prediction is twofold. First, we employ different prediction methods, including (1) the established logit model; (2) other well-known, easy-to-use algorithms, such as discriminant analysis or tree-based procedures; (3) advanced single classifiers, such as artificial neural networks or support vector machines; and (4) powerful off-the-shelf ensembles, such as bagging or boosting (e.g., Lemmens and Croux, 2006). Second, we vary the meta-parameter settings of individual learners. Meta-parameters allow the analyst to adapt a prediction method to a particular modeling task (Hastie et al., 2009). This suggests that a single method will produce somewhat different models if it is invoked with different settings for algorithmic parameters. Table 1 summarizes the classification methods and meta-parameter settings in our model library. Our particular selection of methods and meta-parameters is based on previous churn modeling studies (e.g., Verbeke et al., 2012) and literature recommendations (e.g., Caruana et al., 2004; Partalas et al., 2010).

Classification Method	**Number of models**	**Meta-parameter**	**Candidate Settings**
		Single Classifiers	
Classification and Regression Tree (CART)	6	Min. size of nonterminal nodes Pruning of fully grown tree	10, 100, 1000 Yes, No
Artificial Neural Network (ANN)	162	No. of neurons in hidden layer Regularization factor (weight decay)	1, 2, …, 20 $10^{[-4, -3.5, \ldots, 0]}$
k-Nearest-Neighbor (kNN)	18	Number of nearest neighbors	10, 100, 150, 200, …, 500, 1000, 1500, …4000
Linear Discriminant Analysis (LDA)	20	Covariates considered in the model	Full model, stepwise variable selection with p-values in the range 0.05, 0.1,…, 0.95
Logistic Regression (LogR)	20	Covariates considered in the model	Full model, stepwise variable selection with p-values in the range 0.05, 0.1,…, 0.95
Naive Bayes (NB)	9	Histogram bin size	2, 3, …, 10
Quadratic Discriminant Analysis (QDA)	20	Covariates considered in the model	Full model, stepwise variable selection with p-values in the range 0.05, 0.1,…, 0.95
Regularized Logistic Regression (RLR)	29	Regularization factor	$2^{[-14, -13, \ldots, 14]}$
Support Vector Machine with linear kernel (SVM-Lin)	29	Regularization factor	$2^{[-14, -13, \ldots, 14]}$
Support Vector Machine with Radial Basis Function Kernel (SVM-Rbf)	300	Regularization factor Width of Rbf kernel function	$2^{[-12, -11, \ldots, 12]}$ $2^{[-12, -11, \ldots, -1]}$
		Ensemble Learners	
AdaBoost (AdaB)	11	No. of member classifiers	10, 20, 30, 40, 50, 100, 250, 500, 1000, 1500, 2000
Bagged Decision Trees (BagDT)	11	No. of member classifiers	10, 20, 30, 40, 50, 100, 250, 500, 1000, 1500, 2000
Bagged Neural Networks (BagNN)	5	No. of member classifiers	5, 10, 25, 50, 100
Random Forest (RF)	35	No. of member classifiers No. of covariates randomly selected for node splitting	100, 250, 500, 750, 1000, 1500, 2000 $[0.1, 0.5, 1, 2, 4] \cdot \sqrt{M}$
LogitBoost (LoB)	11	No. of member classifiers	10, 20, 30, 40, 50, 100, 250, 500, 1000, 1500, 2000
Stochastic Gradient Boosting (SGB)	11	No. of member classifiers	10, 20, 30, 40, 50, 100, 250, 500, 1000, 1500, 2000

Table 1. Classification methods and meta-parameter settings employed in the study.

4.3. Candidate Selection

Following Caruana et al. (2004), we initialize candidate selection with finding the best performing individual churn model in our library. To improve performance, we then assess all pairwise combinations of this model and one other model from the library. We select the best-performing size-two ensemble if it outperforms the best individual model. Next, we examine the best-performing ensemble of size three. That is, we assess all combinations of the current size-two ensemble and one other candidate model from the library. The stepwise ensemble growing procedure stops as soon as appending additional members does not improve performance. The candidate selection strategy of Caruana et al. (2004) is able to accommodate any objective function that depends on the estimated churn probabilities. We exploit this feature for our DCES approach. In particular, we organize candidate selection in such a way that it maximizes the lift index. Recall that the lift is directly connected to the profitability of retention campaigns (Neslin et al., 2006). Therefore, by maximizing lift during candidate selection, we devise ensembles that explicitly pursue actual business objectives (i.e., campaign profits) during model building.

Finally, note that assessing alternative model combinations requires auxiliary validation data. That is, we need one set of data to build the candidate models in the library, and a second set of (validation) data to calculate the lift of individual and combined models during candidate selection. We construct these two samples by means of cross-validation because previous research find it superior to alternative regimes (Partalas et al., 2010).

4.4. Forecast Combination

A combination of multiple prediction models occurs during candidate selection and also when the final ensemble is employed to generate churn scores for novel customers. We pool models by averaging over their predictions. More specifically, given that the candidate selection procedure allows models to enter the ensemble multiple times, we effectively compute a weighted average (Caruana et al., 2004). The opportunity to weight base model predictions in the ensemble whenever the data suggest that some members deserve a greater influence on the composite forecast adds to the flexibility of ensemble selection and may increase performance under certain circumstances.

Finally, note that averaging model predictions is feasible only if all models produce forecasts of a common scale. To achieve this, we convert all model predictions into churn probabilities. Specifically, we project model outputs to the interval [0, 1] by means of a logistic link function (Platt, 2000).

5. Data

We examine our research questions in an empirical study related to telecommunications churn. Customer attrition has been well addressed in this industry so that sophisticated variables to predict churn are available (e.g., Kim, 2010). Consequently, it is particularly challenging to outperform conventional churn models on real-world telecommunications data. Table 2 provides a summary of the eight real-life churn datasets used in our study.

The number of covariates to model the binary response variable *churn = yes/no* varies from 20 (*UCI*) to 359 (*EuroOp*). Each data set contains continuous and categorical predictors. Most of the variables in all the data sets are associated with call detail records, customer demographics, contract characteristics, relational information, or billing data. For each data set, we perform several preprocessing operations (e.g. elimination of linear dependency and missing values). We then create two versions of each data set, one for prediction methods that can process categorical data (e.g., tree-based methods) and one for methods such as neural networks that require an additional category encoding (e.g., Crone et al., 2006).

In the latter case, we transform each categorical variable into a set of indicator variables to represent every possible category with one binary variable. Finally, we randomly partition the data sets into an in-sample training set (60%) and a holdout test set (40%). We use the training and testing partition to build and evaluate prediction models, respectively (e.g., Shmueli and Koppius, 2011).

Data Set	Customer Records	Description / Source
Duke 1- 4	12,410 - 93,893	U.S. customers (http://www.fuqua.duke.edu/centers/ccrm/datasets/download.html)
EuroOp	21,143	European telecommunications carrier
KDD09	50,000	European telecommunications carrier (http://www.sigkdd.org/kdd-cup-2009-customer-relationship-prediction)
Operator	47,761	U.S. domestic carrier (Mozer et al., 2000)
UCI	5000	publicly available data set (www.sgi.com/tech/mlc/db)

Table 2. Telecommunication datasets used for the validation of our DCES framework.

6. Results

This section reports our results. First, we compare DCES to previous churn models. Next, we examine whether our specific candidate selection strategy explains the observed performance differences. In accordance with previous literature, we use the top-decile-lift, $L_{.1}$, as performance measure (e.g., Lemmens and Croux, 2006; Risselada et al., 2010).

6.1. RQ1: Does DCES Outperform the Popular Logit Choice Model?

We compare DCES to the best of 20 alternative logit choice models and find that DCES produces higher lift scores on all eight churn data sets (Table 3). Next to this performance indicator we also state the size of the final ensemble as well as the model composition of each ensemble. On the basis of a Wilcoxon signed-rank test, the recommended approach for comparing two classifiers (Demšar, 2006), we conclude that DCES performs significantly better than the logit choice model (S = 0, $p = .008$). We then compute the median of the pairwise differences of the two models' lift scores. This measure is a robust estimate of the expected performance difference between DCES and the logit choice model when working with other data sets (García et al., 2010). Our results suggest that the difference amounts to .185 units in lift. Additionally we state the size and synthesis of the final ensemble in Table 3.

Data Set	DCES	LogR	Percent Improvement	Size Final Ensemble	Final Ensemble Composition
Duke 1	**1.471**	1.330	11%	10	ANN, BagDT, CART, kNN, SGB, SVM-Lin, SVM-Rbf
Duke 2	**1.612**	1.419	14%	4	BagDT, CART, SVM-Rbf
Duke 3	**2.444**	2.159	13%	8	ANN, SVM-Lin
Duke 4	**1.838**	1.500	23%	8	AdaB, ANN, kNN, RF, RLR, SVM-Rbf
EuroOp	**2.622**	2.446	7%	7	AdaB, ANN, CART, RF, SVM-Lin
KDD09	**1.885**	1.837	3%	10	AdaB, ANN, LoB, RF, SVM-Rbf
Operator	**3.770**	3.673	3%	10	ANN, BagNN, LoB, SVM-Rbf
UCI	**6.821**	3.500	95%	2	RF, RLR
		.185	*= Median difference DCES vs. LogR*		

Table 3. Performance of DCES versus the logit choice model in terms of $L_{.1}$ as well as the size of the final ensemble and model composition.

The superior performance of DCES may seem trivial. It is an advanced modeling paradigm and can capitalize on a large library of candidate models when forming the ensemble. However, the logit choice model is still an important benchmark because of its popularity in marketing (e.g., Cui and Curry, 2005).

6.2. RQ2: How Does DCES Perform in Comparison to Advanced Single Classifiers?

A variety of single classifiers have been considered for churn prediction (e.g., Verbeke et al., 2012). Many of these are more advanced than the logit choice model and thus represent a more challenging benchmark. We compare DCES to nine such methods in Table 4, and find that DCES gives the highest lift scores in all comparisons. To confirm the significance of this result, we test the null-hypothesis of equal performance using the Friedman test (Demšar, 2006), and reject it with high confidence (Friedman's $\chi 2 = 43.47$, d.f. = 9, $p < .001$). We then compute the following test statistic for all $k-1$ pairwise comparisons of DCES with one other churn model (García et al., 2010):

$$z_j = \left(R_{ES} - R_j\right) \Big/ \sqrt{\frac{k(k+1)}{6n}}, \qquad (2)$$

where R_{ES} and R_j represent the average rank of DCES and benchmark j, respectively, and n is the number of data sets. We can translate the z_j into a probability (p_j) using the standard normal distribution table. The resulting p-values require further adjustment to control the family-wise error level and ensure an overall significance level of $\alpha = .05$. We use the Hommel procedure for this purpose because it is one of the most powerful approaches available (García et al., 2010). The adjusted p-values (p_j *adj.*) corresponding to the pairwise comparisons indicate that DCES performs significantly better than the single classifiers (Table 4). The last row of Table 4 depicts the improvement of DCES over a benchmark churn model that can be expected when using other data than used in the study. We develop this statistic using the contrast estimation approach of García et al. (2010). The expected differences range from approximately one-quarter to a full unit in $L_{.1}$.

Data Set	DCES	RLR	ANN	SVM-Lin	SVM-Rbf	NB	kNN	QDA	LDA	CART
Duke 1	**1.471**	1.325	1.248	1.317	1.337	1.219	1.276	1.294	1.331	1.120
Duke 2	**1.612**	1.425	1.505	1.422	1.477	1.042	1.371	1.332	1.424	1.116
Duke 3	**2.444**	2.221	2.402	2.107	2.345	1.388	2.138	1.905	2.133	1.942
Duke 4	**1.838**	1.500	1.576	1.523	1.452	1.294	1.446	1.394	1.493	1.513
EuroOp	**2.622**	2.289	2.133	2.456	2.055	1.624	1.908	2.201	2.387	1.272
KDD09	**1.885**	1.823	1.748	1.851	1.213	0.932	1.542	1.707	1.775	1.200
Operator	**3.770**	1.363	3.520	1.628	3.088	1.085	3.450	3.269	3.673	2.379
UCI	**6.821**	3.143	5.893	2.786	5.857	1.000	4.321	3.643	3.179	4.429
Avg. rank	1.000	5.125	3.750	5.250	4.875	9.750	6.375	6.750	4.500	7.625
p_j *adj.*		.0125	.050	0.01	.017	.006	.008	.007	.025	.006
Contrast DCES vs. classifier j		.3278	.2270	.3177	.3331	.9028	.3786	.4047	.2835	.6127

Table 4. Performance of DCES versus single classifiers in terms of $L_{.1}$.

6.3. RQ3: Can DCES Beat Standard Ensemble Learners?

Previous studies suggest that standard ensemble algorithms represent the most challenging benchmark in churn modeling (e.g., Lemmens and Croux, 2006; Risselada et al., 2010). We compare DCES with six state-of-the-art ensembles, including stochastic gradient boosting, which was the best-performing method in the Duke/NCR Teradata Churn Modeling Tournament (Neslin et al., 2006).

Data Set	DCES	BagDT	BagNN	RF	AdaB	SGB	LoB
Duke 1	**1.471**	1.457	1.382	1.466	1.406	1.435	1.415
Duke 2	**1.612**	1.590	1.495	1.601	1.565	1.554	1.560
Duke 3	**2.444**	2.392	2.423	2.387	2.330	2.247	2.278
Duke 4	**1.838**	1.811	1.651	1.800	1.671	1.760	1.728
EuroOp	2.622	2.407	2.368	2.358	2.417	2.642	**2.661**
KDD09	1.885	1.542	1.775	1.707	1.864	1.878	**1.899**
Operator	3.770	3.172	3.812	3.575	**3.895**	3.631	3.700
UCI	**6.821**	6.750	5.964	6.786	4.214	4.214	4.571
Avg. rank	1.625	4.125	5.000	4.000	4.563	4.688	4.000
p_j *adj.*		.017	.008	.025	.013	.010	.050
Contrast DCES vs. ensemble j		.0506	.1131	.0451	.0871	.0761	.0710

Table 5. Performance of DCES versus standard ensembles in terms of L.1.

Table 5 illustrates that DCES achieves a much better performance (e.g., lower average rank) than RF and LogitBoost (LoB), the two second best models in the comparison (1.625 vs. 4.000). Using the Friedman test, we reject the null hypothesis of equal performance (Friedman's $\chi_2 = 12.76$, d.f. = 6, $p = .0470$). Furthermore, Hommel's procedure rejects all pairwise hypotheses of equal performance between DCES and one other standard ensemble at $\alpha = .05$ for the adjusted p-values in Table 5. Given that the ensemble benchmarks have shown excellent performance in previous research (e.g., Ha et al., 2005; Lemmens and Croux, 2006; Verbeke et al., 2012), outperforming these methods with significant margin provides strong evidence for the effectiveness of DCES. However, the advantage in terms of expected gains in lift (last row of Table 5) is smaller than in previous comparisons. In this sense, Table 5 confirms the competitiveness of standard ensemble methods.

6.4. Does Our Lift-Based Modelling Philosophy Explain the Performance of DCES?

It is important to understand which factors explain the success of DCES, and to confirm that its appealing performance is a consequence of our choice to incorporate the lift measure into the model building process in particular. Three main characteristics distinguish DCES from previous churn models: (1) the availability of a large library of candidate models, (2) the practice to average multiple models' predictions, and (3) the lift-maximizing ensemble selection strategy. In the following, we explore the individual importance of these factors to obtain a clear view on their relative merits.

6.4.1. Library Size

DCES has access to a library of candidate models. To test whether DCES requires a large model library and to which extent smaller libraries are still effective, we randomly delete 2% of the candidate models from the library, create an ensemble using DCES, and assess its performance in terms of *L*.1. We repeat this procedure 50 times, each time reducing the size of the library by two percent. Figure 1 depicts the corresponding development of DCES performance as well as the lift-scores of the logit model (LogR), ANN, and the LoB ensemble for comparative purpose. In general, Figure 1 reveals that DCES is robust toward a random elimination of candidate models. Even small libraries of approximately 50 models suffice to perform well. In particular, DCES is consistently better than LogR and ANN. Furthermore, reducing the library size never decreases the performance of DCES below the LoB level in settings where DCES has originally outperformed LoB. In view of these results, we conclude that the size of the model library does not explain the success of DCES.

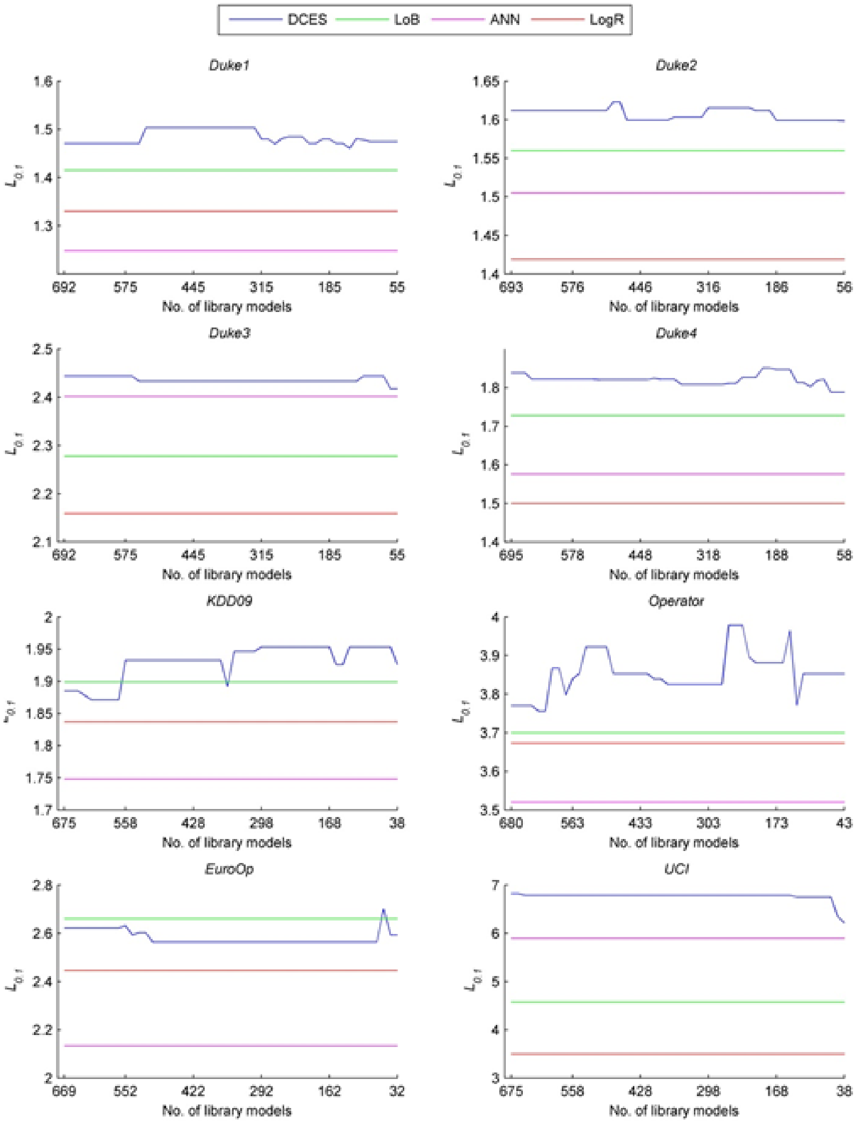

Figure 1. *Development of DCES performance when repetitively removing 2% (of the original library size) randomly selected candidate models for 50 iterations.*

6.4.2. Forecast Averaging

A second characteristic of DCES is that it develops a composite forecasting. To clarify the degree to which this feature explains the success of DCES, we compare DCES to four popular forecast combination approaches (e.g., Armstrong, 2001): (1) a simple average (SAvg); (2) a weighted average (WAvg); (3) a trimmed average (TAvg) which discards the *n*% most extreme churn predictions and (4) a weighted average resulting from regressing the binary churn indicator variable on the library models' predictions (RAvg). We calculate the weight of library model on the basis of its performance on the validation sample. Similarly, TAvg and RAvg employ the validation sample to select the trimming fraction *n* from the interval [.5, .1, …, .95] and to build the regression model, respectively.

With the exception of the *Operator* data set, we find that the average-based combination mechanisms perform not as well as DCES (Table 6). In particular, using the statistical testing framework elaborated above, we may conclude that DCES performs significantly better than all average-based competitors but WAvg. Given the estimated contrasts (last row of Table 6), we expect that DCES improves lift by .08 to .99 points on average. Moreover, comparing Table 5 and Table 6, we find that the average-based combination schemes do not improve over the standard ensemble learners which are already well-established in churn prediction. This is noteworthy because DCES operates similar to WAvg and RAvg in that it also forms a weighted average over library model predictors. Despite this similarity, DCES outperforms these competitors. Overall, these results suggest that forecast averaging alone cannot be the reason why DCES performs well.

Data Set	**DCES**	**SAvg**	**WAvg**	**TAvg**	**RAvg**
Duke 1	**1.471**	1.382	1.382	0.941	1.326
Duke 2	**1.612**	1.566	1.568	1.068	1.424
Duke 3	**2.444**	2.361	2.366	1.134	2.195
Duke 4	**1.838**	1.715	1.718	1.077	1.498
EuroOp	**2.622**	2.553	2.553	0.969	0.929
KDD09	**1.885**	1.871	1.878	1.254	1.158
Operator	3.770	3.965	**3.979**	0.543	1.113
UCI	**6.821**	6.143	6.357	1.464	0.179
Avg. rank	1.250	2.750	2.000	4.625	4.375
p_j *adj.*		.025	.050	.013	.017
Contrast ES vs. Avg j		.0802	.0763	.9987	.5633

Table 6. Performance of DCES versus average-based forecast combination in terms of $L_{.1}$.

6.4.3. Lift-Maximizing Candidate Selection

Having ruled out the influence of the library size and the model averaging, we hypothesize that the success of DCES comes mainly from our lift-maximizing candidate selection strategy. Theory suggests that the prosperity of any ensemble is related to the strength and diversity of its members (e.g., Kuncheva, 2004). These goals conflict because perfect models that discriminate between switchers and stayers with maximal accuracy must be perfectly correlated and thus lack diversity. In view of the observed results, we suspect that the appealing performance of DCES stems from its specific candidate selection strategy achieving a better balance between strength and diversity.

To test this, we perform a kappa-lift analysis (Margineantu and Dietterich, 1997). In particular, given an ensemble of n members, we first compute kappa for all $(n \times [n-1])/2$ possible pairs of members and the mean lift score for all possible pairs of ensemble members. This allows us to depict the relationship between strength (i.e., lift) and diversity (i.e., kappa) in a scatterplot (see Figure 2).

By this it becomes obvious that DCES leads to parsimonious ensembles, which normally embrace substantially fewer members than the best LoB ensembles. This is appealing because smaller ensembles consume less memory and predict at higher speeds (e.g., Margineantu and Dietterich, 1997). Furthermore, except for *KDD09* the pairwise lift scores of ensemble members are higher when using DCES. With respect to diversity, except for *Duke 1* and *Duke 3* we observe no trend of ensemble members being less diverse when pursuing a decision-centric modeling strategy. This suggests that differences in diversity between DCES and LoB are not systematic. Therefore, this kappa-lift analysis supports our proposition that the success of DCES is mainly due to maximizing lift during member selection.

By choosing candidate models with high lift, the final ensemble includes only members that perform well in terms of lift. The reason DCES achieves a better balance between strength and diversity in our churn context is precisely that it is able to concentrate on the "right" measure of strength. Standard ensembles strategies also balance strength and diversity. However, their notion of strength is different, internally fixed by the underlying learning algorithm and agnostic of application characteristics. The ensembles resulting from LoB in kappa-lift space exhibit comparable degrees of diversity but at lower levels of strength. This does not mean that DCES is a better modeling approach in general but a more flexible approach that facilitates governing member selection toward arbitrary performance measures. This feature is particularly valuable in applications with some discrepancy between accuracy indicators that are typically incorporated in standard prediction methods and performance measures that matter from a business perspective. Churn prediction is such an application and aims at models with high lift. DCES takes this objective into account, and this is why it outperforms alternative approaches.

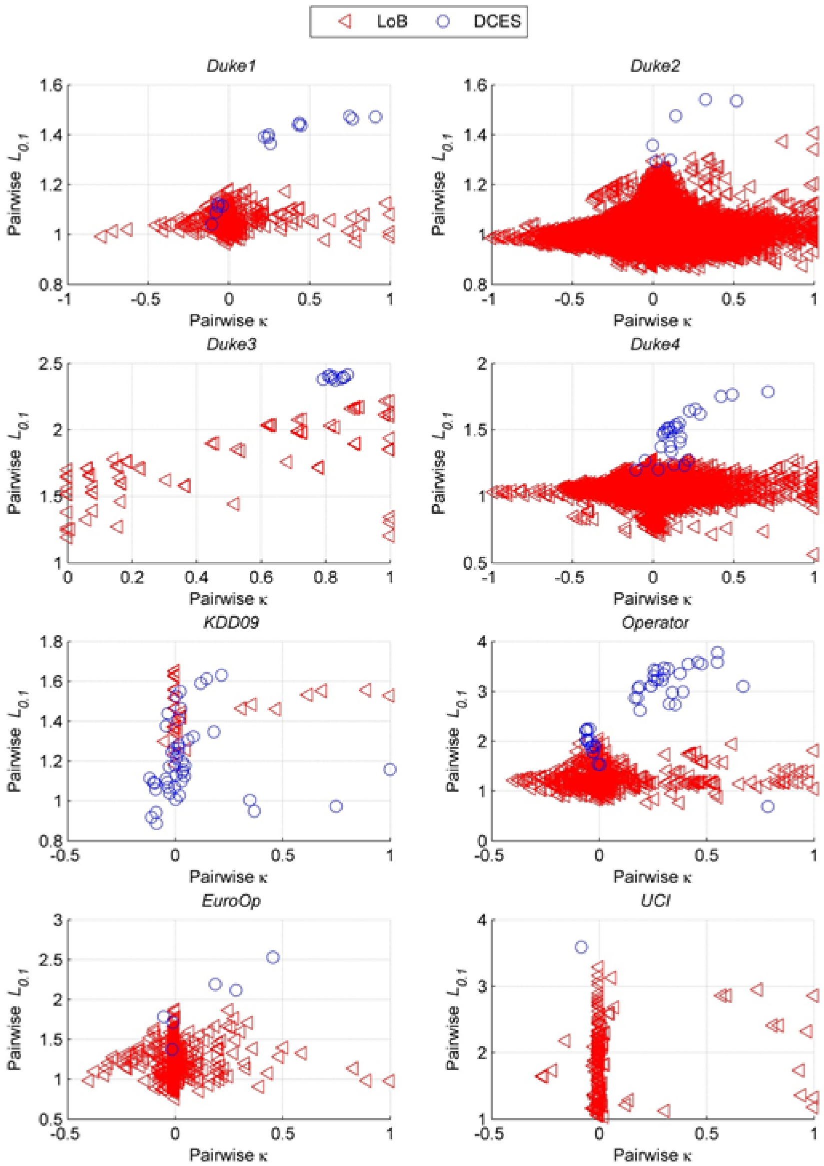

Figure 2. Kappa-lift analysis of the strength and diversity of DCES and LoB ensemble member.

7. Discussion

We set out to develop a framework for decision-centric churn modeling and to test its effectiveness in a large-scale empirical study by comparing our approach to several previous well established modeling approaches (e.g., logit choice model). We find that DCES performs significantly better than any of these benchmarks. Although DCES can benefit from large model libraries in our study, a sensitivity analysis reveals that the number of candidate models is not a key success factor. The unique advantage of DCES stems from the opportunity to organize the model choice process in a way that reflects actual business objectives. Building the ensemble model so as to maximize lift, DCES concentrates on the performance criterion that matters from a business standpoint. We find that this facilitates to balance strength and diversity, the key determinants of ensemble success.

Possible explanations why the DCES approach has not been considered in previous work is that maximizing a discontinuous function such as lift during model fitting is highly challenging from a mathematical point of view. However, a more important reason is that model fitting is an induction problem. Even if we can overcome mathematical obstacles, approaching a statistical problem exclusively from a business angle may not be the right approach after all. A conceptual advantage of our DCES framework is that it unifies these two worlds. It leverages established statistical methods for building the candidate library and then shifts attention to the business perspective when finding the subset of models most suitable for solving the decision problem.

7.1. Implications

Our results have several implications for the science and practice of churn management. First, the finding that the new ensemble selection approach significantly outperforms what is considered the state-of-the-art emphasizes that exploring novel ways to anticipate churn and developing novel modeling frameworks is a fruitful avenue of research. It is still possible to improve on the best models known today, identify likely churners with greater accuracy, and eventually increase the effectiveness of churn management activities.

Second, it is feasible and effective to consider business performance measures when building a churn model. Unlike previous approaches, DCES takes marketing objectives into account. This is more aligned with how managers make decisions and increases the model's fit for the ultimate decision support task. In a churn context, the lift measure captures typical business objectives. Our results confirm the effectiveness to introduce this notion of performance into the model building process.

Third, analysts often test alternative approaches before deploying a final churn model. Such alternatives may originate from exploring different prediction methods and/or from experimenting with different sets of customers. The standard approach is then to pick the single "best" model and discard all the others. Our results suggest that an appropriately chosen combination of some of these alternative models will increase model performance. This selection and combination step is an excellent opportunity to introduce business objectives into the modeling process.

From a managerial perspective, a key question is to what extent better churn models add to the bottom line. Research has shown that customer retention is an important determinant of firm performance (e.g., Gupta and Zeithaml, 2006). Churn prediction aims at targeting retention programs to possible churners and thus supports customer retention. This suggests that an indirect link between accurate churn predictions and firm performance exists. Neslin et al. (2006) examine the profit impact of churn modeling in more detail and quantify a per-customer profit increase of \$1.71 per unit change in lift. We find that the expected improvement of DCES over previous churn models is .276 lift units. This suggests that a company can expect an increase in per-customer profits of \$.47 (\$1.71 × .276) when adopting our DCES approach. Depending on the size of the company a \$.47 increase in per-customer profits can easily amount to changes in profit in the hundreds of thousands of dollars.

Another advantage of DCES is that it requires little human intervention. Modeling tasks typically carried out by the analyst include testing and transformation of covariates and prediction methods. With DCES, it is only necessary to incorporate the candidate models that represent the choice alternatives into the library. The selection strategy will then pick the most beneficial model combination in a decision-centric manner. This frees marketers from laborious, repetitive modeling tasks and opens up valuable resources.

7.2. Avenues for Further Research

Our study suggests several directions for further research. First, DCES works well for predictive modeling but does not allow an interpretation of how customer characteristics influence the estimated churn scores. Since marketers and managers require comprehensible and understandable models it is important to develop procedures that clarify how covariates influence DCES predictions and what are the main drivers of customer churn. Second, all our data sets represent a snapshot drawn at a given point of time. However, churn is a dynamic phenomenon and the causes for defection change over time. It would thus be interesting to explore the potential of DCES in a longitudinal setting.

Third, it is important to validate the appropriateness of DCES in marketing applications other than churn modeling such as, e.g. scoring new product acceptance or estimating direct mail response. The opportunities to account for business objectives and constraints in the model-building process extend to these settings. Reproducing our results and confirming the effectiveness of a decision-centric modeling philosophy in other marketing applications would thus be a particularly fruitful research avenue.

References

Armstrong, J. S. (2001). *Combining Forecasts. Principles of Forecasting: A Handbook for Researchers and Practitioners*. Ed. by J. S. Armstrong. Boston: Kluwer, pp. 417-439.

Bensinger, G. and S. Tibken (2012). *T-Mobile Struggles to Stem Customer Losses*. URL: http://online.wsj.com/articles/SB10001424052702304203604577396393221489630 (visited 11/26/2014).

Bhattacharya, C. B. (1998). "When customers are members: Customer retention in paid membership contexts." *Journal of the Academy of Marketing Science* 26 (1), 31–44.

Bolton, R. N. (1998). "A dynamic model of the duration of the customer's relationship with a continuous service provider: The role of satisfaction." *Marketing Science* 17 (1), 45–65.

Burez, J. and D. Van den Poel (2007). "CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services." *Expert Systems with Applications* 32 (2), 277–288.

Capraro, A. J., Broniarczyk, S. and R. K. Srivastava (2003). "Factors influencing the likelihood of customer defection: The role of consumer knowledge." *Journal of the Academy of Marketing Science* 31 (2), 164–175.

Caruana, R., Niculescu-Mizil, A., Crew, G. and A. Ksikes (2004). "Ensemble Selection from Libraries of Models." In: *Proceedings of the 21st International Conference On Machine Learning*. Ed. by C. E. Brodley. ACM. New York, pp. 18–25.

classification performance of customer churn prediction models." *IEEE Transactions on Knowledge and Data Engineering* 25, 961–973.

Colgate, M. R. and P. J. Danaher (2000). "Implementing a customer relationship strategy: The asymmetric impact of poor versus excellent execution." *Journal of the Academy of Marketing Science* 28 (3), 375–387.

Crone, S. F., Lessmann, S. and R. Stahlbock (2006). "The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing." *European Journal of Operational Research* 173 (3), 781–800.

Cui, D. and D. Curry (2005). "Prediction in marketing using the support vector machine." *Marketing Science* 24 (4), 595–615.

Demšar, J. (2006). "Statistical comparisons of classifiers over multiple data sets." *Journal of Machine Learning Research* 7, 1–30.

Fader, P. S. and B. G. S. Hardie (2010). "Customer-base valuation in a contractual setting: The perils of ignoring heterogeneity." *Marketing Science* 29 (1), 85–93.

Ganesh, J., Arnold, M. J. and K. E. Reynolds (2000). "Understanding the customer base of service providers: An examination of the differences between switchers and stayers." *Journal of Marketing* 64 (3), 65–87.

García, S., Fernández, A., Luengo, J. and F. Herrera (2010). "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power." *Information Sciences* 180 (10), 2044–2064.

Gupta, S. and V. Zeithaml (2006). "Customer metrics and their impact on financial performance." *Marketing Science* 25 (6), 718–739.

Gustafsson, A., Johnson, M. D. and I. Roos (2005). "The effects of customer satisfaction, relationship commitment dimensions, and triggers on customer retention." *Journal of Marketing* 69 (4), 210–218.

Ha, K., Cho, S. and D. MacLachlan (2005). "Response models based on bagging neural networks." *Journal of Interactive Marketing* 19 (1), 17–30.

Hastie, T., Tibshirani, R. and J. H. Friedman (2009). *The Elements of Statistical Learning*. New York: Springer.

Kim, G. (2010). *AT&T Churn Rate Offers Lesson*. URL: http://technews.tmcnet.com/voice-quality/topics/phone-service/articles/93062-att-churn-rate-offers-lesson.htm (visited 11/26/2014).

Kim, Y. S., Lee, H., and J. D. Johnson (2013). "Churn management optimization with controllable marketing variables and associated management costs." *Expert Systems with Applications* 40 (6), 2198–2207.

Kopalle, P. K., Sun, Y., Neslin, S.A., Sun, B. and V. Swaminathan (2012). "The joint sales impact of frequency reward and customer tier components of loyalty programs." *Marketing Science* 31 (2), 216–235.

Kuncheva, L. I. (2004). *Combining Pattern Classifiers Methods and Algorithms*. Hoboken: Wiley.

Lemmens, A. and C. Croux (2006). "Bagging and boosting classification trees to predict churn." *Journal of Marketing Research* 43 (2), 276–286.

Lessmann, S., and S. Voß (2010). "Customer-centric decision support: A benchmarking study of novel versus established classification models." *Business & Information Systems Engineering* 2, 79–93.

Lewis, M. (2004). "The influence of loyalty programs and short-term promotions on customer retention." *Journal of Marketing Research* 41 (3), 281–292.

Lilien, G. L. (2011). "Bridging the academic–practitioner divide in marketing decision models." *Journal of Marketing* 75 (4), 196–210.

Ling, C. X. and C. Li (1998). "Data Mining for Direct Marketing: Problems and Solutions." In: *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*. Eds. by R. Agrawal, P. E. Stolorz, G. Piatetsky-Shapiro. AAAI Press. Menlo Park, pp. 73–79.

Loterman, G., Brown, I., Martens, D., Mues, C., and B. Baesens (2012). "Benchmarking regression algorithms for loss given default modeling." *International Journal of Forecasting* 28, 161–170.
Margineantu, D. D. and T. G. Dietterich (1997). "Pruning Adaptive Boosting." In: Proceedings of the 14th International Conference on Machine Learning. Ed. by D. H. Fisher. Morgan Kaufmann. San Fransisco, pp. 211–218.
Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E. and H. Kaushansky (2000). "Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry." *IEEE Transactions on Neural Networks* 11 (3), 690–696.
Musalem, A. and Y. V. Joshi (2009). "How much should you invest in each customer relationship? A competitive strategic approach." *Marketing Science* 28 (3), 555–565.
Neslin, S. A., Gupta, S., Kamakura, W., Lu, J. and C. H. Mason (2006). "Defection detection: Measuring and understanding the predictive accuracy of customer churn models." *Journal of Marketing Research* 43 (2), 204–211.
Partalas, I., Tsoumakas, G. and I. Vlahavas (2010). "An ensemble uncertainty aware measure for directed hill climbing ensemble pruning." *Machine Learning* 81 (3), 257–282.
Platt, J. C. (2000). *Probabilities for Support Vector Machines. Advances in Large Margin Classifiers*. Eds. by A. Smola, P. Bartlett, B. Schölkopf, D. Schuurmans. Cambridge: MIT Press, 61–74.
Reichheld, F. F. (1996). "Learning from customer defections." *Havard Business Review* 74 (2), 56–69.
Risselada, H., Verhoef, P. C. and T. H. A. Bijmolt (2010). "Staying power of churn prediction models." *Journal of Interactive Marketing* 24 (3), 198–208.
Schweidel, D. A., Fader, P. S. and E. T. Bradlow (2008). "Understanding service retention within and across cohorts using limited information." *Journal of Marketing* 72 (1), 82–94.
Shmueli, G. and O. R. Koppius (2011). "Predictive analytics in information systems research." *MIS Quarterly* 35 (3), 553–572.
Thomas, J. S., Blattberg, R. and E. Fox (2004). "Recapturing lost customers." *Journal of Marketing Research* 41 (1), 31–56.
Verbeke, W., Dejaeger, K., Martens, D., Hur, J. and B. Baesens (2012). "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach." *European Journal of Operational Research* 218 (1), 211–229.
Verbraken, T., Verbeke, W., and B. Baesens (2012). "A novel profit maximizing metric for measuring classification performance of customer churn prediction models." In: *IEEE Transactions on Knowledge and Data Engineering* 25 (5), 961–973.
Verhoef, P. C. (2003). "Understanding the effect of customer relationship management efforts on customer retention and customer share development." *Journal of Marketing* 67 (4), 30–45.
Zeithaml, V. A., Berry, L. L. and A. Parasuraman (1996). "The behavioral consequences of service quality." *Journal of Marketing* 60 (2), 31–46.

ARTICLE 8:

CHANGING PERSPECTIVES: USING GRAPH METRICS TO PREDICT PURCHASE PROBABILITIES[10]

Authors: Annika Baumann, Humboldt-Universität zu Berlin, Berlin, Germany
Johannes Haupt, Humboldt-Universität zu Berlin, Berlin, Germany
Fabian Gebert, Akanoo GmbH, Hamburg, Germany
Stefan Lessmann, Humboldt-Universität zu Berlin, Berlin, Germany

Journal: Expert Systems with Applications

Publiation year: 2018

Abstract

The prediction of online user behavior (next clicks, repeat visits, purchases, etc.) is a well-studied subject in research. Prediction models typically rely on clickstream data that is captured during the visit of a website and embodies user agent-, path-, time- and basket-related information. The aim of this paper is to propose an alternative approach to extract auxiliary information from the website navigation graph of individual users and to test the predictive power of this information. Using two real-world large datasets of online retailers we develop an approach to construct within session graphs from clickstream data and demonstrate the relevance of corresponding graph metrics to predict purchases.

Keywords: *Predictive Analytics, Clickstream Data, User Graph, Graph Metrics*

[10] This article is provided with kind permission from Elsevier. The original version is available at: https://doi.org/10.1016/j.eswa.2017.10.046.

1. Introduction

The e-commerce sector is responsible for a substantial fraction of firm revenues. Annual turnover was 1,336 billion US dollars in 2014 and is predicted to have reached 2,050 billion US dollars in 2016 (Statista, 2016a). However, given that growth rates are expected to decline in the future (Statista, 2016b), e-commerce shops need to find ways to defend market shares in an increasingly competitive environment. One strategy is to increase purchasing amounts and/or frequencies of existing customers. Important determinants of (re-)purchase intention in online shopping are trust, service quality (Hong and Kim, 2012) and user satisfaction during the online shopping process (Lee et al., 2009). To offer a richer user experience and increase visitors' (re-)purchase intentions, understanding customer online behavior is crucial (e.g., Pai et al., 2014). To gain such insight and to anticipate user actions, the analysis of clickstream data has been widely adopted in the literature (e.g., Van den Poel and Buckinx, 2005; Park and Park, 2016).

However, previous work in the field has not examined the potential of graph theory to gather auxiliary information from clickstream data and increase the accuracy of behavior prediction models. Graphs are a methodological approach originating from network theory. They consist of nodes and edges, which connect nodes. Graph-based approaches have been used in various fields and have been proven to be helpful for various tasks, for example to predict connections in the social networking context (He et al., 2015), to detect money laundering activities (Colladon and Remondi, 2017), for personalized recommendations (Shams and Haratizadeh, 2017) and for customer churn prevention (Óskarsdóttir et al., 2017). Given the success of graph-based predictors in these and other applications, the objective of our paper is to test their potential for online behavior prediction based upon clickstream data.

We contribute to literature as follows: First, we propose an approach to derive graphs from user sessions based on clickstream data. Second, we calculate graph metrics and examine their pairwise dependency in terms of correlation. Third, we assess how they perform as a means to predict customer behavior in online contexts.

The remainder of the paper is structured as follows. First, we give an overview on relevant literature to clarify the research gap the paper strives to close. Afterwards, we present our methodology and how we derive clickstream graphs in particular. We then summarize the resulting data, before presenting empirical results. Lastly, we summarize our findings.

2. Related Work

Much literature considers the use of clickstream data for customer online behavior prediction. Prediction targets range from conversions in purchase prediction (Van den Poel and Buckinx, 2005), whether visitors redeem incentives (Pai et al., 2014) or complete specific tasks such as putting an item into a basket (Sismeiro and Bucklin, 2004; Kalczynski et al., 2006), over navigational behavior prediction (e.g., the next web path access; Montgomery et al., 2004) to classifying visitors into interest groups such as whether a user's site visiting intention is informational or transactional (Moe, 2003).

Table 1 summarizes related work, which we categorize according to the target of prediction into navigational behavior (NB), user classification (UC) and conversion (PC) prediction, where PC is the prevailing target in prior work (i.e., 23 out of 34 studies fall in this category).

Table 1 also shows the types of features (i.e., covariates) which the studies employ for predictive modelling. In particular, we categorize the features into six groups. All categories except demographics are based on clickstream data. The first three groups – time, page and monetary – draw inspiration from the well-known concept of recency, frequency and monetary value analysis (Zhang et al., 2015).

Recency and frequency consist of aspects such as time on page and last website visit (*Time*), whereas monetary comprises historical purchase behavior derived from preceding clickstream sessions and current basket information (*Monetary*). Frequency refers to the path traversal and categories of pages visited, counting how often each page has been visited (*Page*). In addition, we consider behavior related variables (*Page Interaction*), such as basket interaction, click on page and scroll on page events to capture user-centered feature categories, which revolve around behavioral aspects besides the website path that a user traverses. The feature category *Demographics* consists of variables that capture user characteristics, such as gender and geographic-related information, which are not related to the website itself and thus not part of clickstream data. The last category captures studies which use graphs as a tool to derive features for their models used (*Graph / Similarity*).

Only four of the 34 studies, which base their analysis on clickstream, use a graph-based approach. In view of Table 1 it becomes evident that combining predictive modelling with graph-based features has been rare so far. Byeon (2013) generates a bi-partite graph (i.e. a graph with two different types of nodes) for each user session, where the nodes represent a specific webpage a user visits during her session and the category to which the webpage belongs, respectively. Each graph facilitates the calculation of summary statistics (i.e., density), which Byeon (2013) employs to predict whether a user session leads to a purchase using a logistic regression classifier. In comparison to classical clickstream features (e.g. total number of clicks, total visit time), the graph density feature provides encouraging results, suggesting that it is a good predictor of purchase intention.

Kalczynski et al. (2006) predict whether a specific website task has been completed successfully. To achieve this, they focus on navigational complexity. The authors construct an experimental setup where users are asked to browse a website to conduct an artificial purchase. The website data is based on five different datasets which they use to derive graphs from user journeys on a website. The authors then calculate a set of graph measures, some of which are based on specific website characteristics. Finally, they employ logistic regression to predict online task completion and conversion in particular.

Other approaches aiming at user classification do so via clustering using graph-based approaches to be able to build similarity graphs, connecting users which behave similarly on websites. Banerjee and Ghosh (2001) use clickstream data to create a similarity graph, which connects users who display similar website usage behavior. First, they select pair-wise user sessions and compare them in terms of path and time dimensions to derive a similarity score. They then construct a weighted graph with nodes representing users which are pairwise connected once the weight reaches a specific threshold. The weight represents the similarity between two users. The similarity graph serves as input to a graph-based clustering method to derive user groups.

A similar approach is applied by Gündüz and Özsu (2003) who construct a similarity graph to apply a graph-based clustering method. Their graph is based on path and time aspects associated with a user journey on a website. The aim of graph construction and clustering is to predict the next website request.

The review of related work suggests that a comprehensive study, which systematically assesses the predictive value of a broad set of graph metrics is lacking. Building on the work of Byeon (2013) and Kalyznski et al. (2015) to predict purchase intention/conversion, we contribute toward closing this research gap in that we i) develop a way to derive a graph from clickstream data, ii) consider a much richer set of graph metrics, iii) employ real-life data, and iv) use state-of-the-art prediction algorithms (random forest, gradient boosting machine) alongside logistic regression.

Reference	Dependent Variable	Feature Category					
		Page	Time	Monetary	Page Interaction	Demo-graphics	Graph / Similarity
Anitha 2010	NB	x					
Antonellis et al. 2009	UC	x					
Banerjee and Ghosh 2001	UC	x	x				x
Berka and Labsky 2007	NB	x					
Byeon 2013	PC	x	x				x
Chan et al. 2014	PC	x			x	x	
Girija and Kavitha 2013	NB	x					
Iwanaga et al. 2016	PC	x	x				
Jiang et al. 2012	PC	x	x				
Kalczynski et al. 2006	PC	x					x
Lee et al. 2010	PC	x	x				
Lu et al. 2005	UC	x					
Moe 2003	UC	x	x				
Moe and Fader 2004	PC		x	x			
Moe et al. 2002	PC	x	x				
Montgomery et al. 2004	NB	x					
Gündüz and Özsu 2003	NB	x	x				x
Padmanabhan et al. 2006	PC / Revisit	x	x	x		x	
Pai et al. 2014	UC	x	x				
Panagiotelis et al. 2014	PC	x	x	x			
Park et al. 2008	UC	x					
Park and Park 2016	PC	x					
Pitman and Zanker 2010	PC	x			x		
Sarwar et al. 2015	PC	x	x	x			
Sato and Asahi 2012	PC (day)	x		x			
Senecal et al. 2014	UC	x	x		x		
Sismeiro and Bucklin 2004	PC	x	x		x		
Stange and Funk 2015	PC	x	x	x	x		
Suh et al. 2004	PC	x	x				
Van den Poel and Buckinx 2005	PC	x	x	x		x	
Vroomen et al. 2005	PC		x	x	x	x	
Wu et al. 2005	PC	x					
Zhao et al. 2016	PC	x	x	x			
Zheng et al. 2003	PC	x	x	x		x	

Table 1. Overview of feature categories used in research (NB: Navigational Behavior, UC: User Classification, PC: Conversion).

3. Methodology

The following sections explain our approaches to create clickstream graphs and derive corresponding graph metrics as input for predictive modeling.

3.1. Clickstream and Graph Construction

A clickstream is defined as the path which a website user traverses when visiting a number of websites (Bucklin et al., 2002) and consists of sessions each of which represent a single visit of a user on a website. Each session consists of an arbitrary number of page views, which are the webpages the user visits during a session. Specific behavior can be performed on a webpage such as click, scroll and basket events. Furthermore, single webpages are visited for a specific amount of time. So far, the representation of clickstream in the form of a graph has been established mainly for visualization purposes (e.g., Kitts et al., 2002). Using clickstream data, we construct a graph for each session of a user to be able to derive covariates for purchase prediction. In general, a graph $G = (V, E)$ consists of a set of nodes V which are pair-wise connected via edges E. The edges are either directed or undirected. Each graph can be represented as a $n \times n$ adjacency matrix where elements a_{ij} are set to one if node n_i and n_j are connected, and zero otherwise.

The user session graphs applied in this paper are constructed in the following way: Each node represents a specific website a user has visited during the session. For each page view, we create a new node if it does not already exist in the session graph (i.e., the user has not visited the page before). We connect two nodes with an edge (i.e., between two pages), if the user visits them successively. The edges are directed to capture the specific order in which webpages are visited. Due to incremental node insertion, the session graph grows successively during the users' journey on the website. This technique is known as "clipping at every click" (Van der Meer et al., 2000), meaning that, for every page view, we calculate a new graph and its underlying graph metrics to capture the user sessions' characteristics in an incremental manner. Figure 1 shows an example of this approach where a session of a user is represented as a graph structure which is updated at every page view, i.e., every webpage the user visits during her journey on the website. Furthermore, as an example the incremental calculation of a graph metric (i.e., average in-degree, which is the average number of edges converging to a node) is shown, which is re-calculated at each page view of a user.

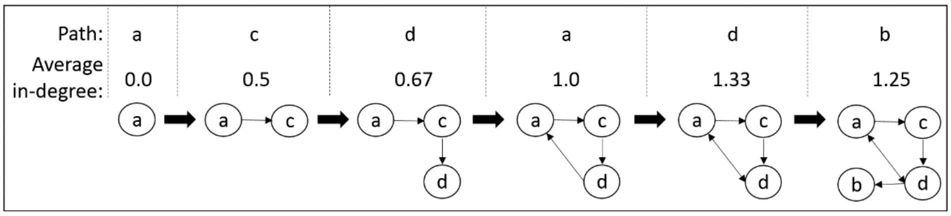

Figure 1. Example of a graph inference of a user session based on clickstream data.

Based upon Kalczynski et al. (2006), we assume that the structure of a graph represents user behavior, which motivates examining the potential of graph metrics to predict conversion. More specifically, a graph based on clickstream data grounds on the explicit user click behavior and captures the path traversal of a user on a website. This behavior is a direct result of the user's goal in visiting the website and changes observably with user intention. To illustrate this, exemplary session graphs for three different types of user behavior are shown in Figure 2.

The left graph shows a direct clickstream path, traversing from one page to another without returning behavior. This indicates a high degree of goal-orientation, which can be associated with both informational (the sought-after piece of information was found) or transactional (the desired product was bought) behavior. The middle graph illustrates a customer looking at various products of two types of product categories before deciding which is of further interest. This is a typical comparison behavior. The right-most graph depicts broad browsing behavior. It contains several central nodes signifying, for example, the overview pages of search results to which the user returns after looking at a sequence of products. These and other graph structures are captured by the graph metrics we apply and reduce to input suitable for predictive modelling.

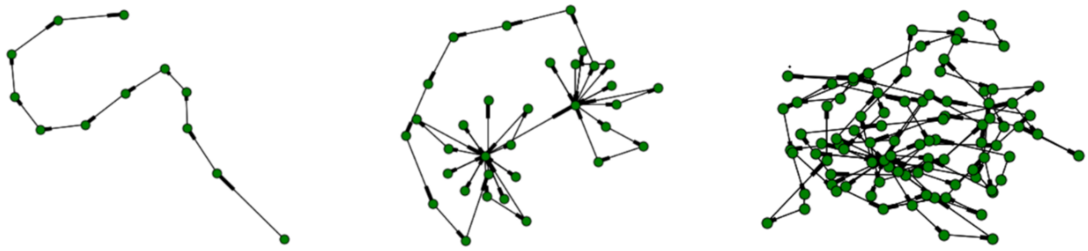

Figure 2. Graph visualizations of user sessions representing different types of user behavior.

3.2. Selected Graph Metrics

Several metrics have been established in graph theory to represent the characteristics of a graph. These metrics focus on a specific node, the n-hop neighborhood or the whole graph. Given our objective to clarify the relative predictiveness of different graph metrics in combination with the scarcity of prior work on graph-based online behavior prediction, it is not clear which graph metrics are the most informative. Therefore, we use the Python framework NetworkX (Hagberg et al., 2008) to create a large number of alternative metrics and compare their relative predictiveness empirically. To that end, we use a set of 23 graph measures in total (see Table 2). Metrics focusing on characteristics of single nodes are computed for all nodes in the graph and then averaged. We concentrate on structural, centrality- and distance-based metrics since they are able to describe the relative importance of graph elements in the network and the general structure of a graph. This in turn might be indicative of the users' intention behind the website visit.

Structural measures describe the general construction of a graph. The most basic concepts are the total number of nodes *N,* capturing the total number of unique webpages which a user visits, and the total number of edges *M,* being the number of directed and unique path traversals from one page to another. The number of circles and self-loops in a graph accounts for switching behavior of a user returning to a previously visited page or the same page, respectively. Related to these measures is flow hierarchy which states the proportion of edges not being part of a cycle. Transitivity indicates whether a user resolves around a specific subset of webpages such as switching between different products to choose from. Unlike transitivity, which focuses on the neighborhood of a node, density can be used as an additional measure of purposefulness of a session, since small values indicate a step-by-step list of pages without circularity or jump-backs hinting at more goal-oriented user behavior. Last, high values of the metric average node connectivity signal an interwoven connection between a considerable number of nodes and therefore a non-structured browsing behavior of users.

Distance measures relate to the broadness of a graph structure. The average shortest path length and the related metrics eccentricity, diameter, radius, center and periphery give an indication how diverse the user traverses through the website. Their intuitive interpretation is that for high values of, e.g., the average shortest path length, the user rather looks at unique webpages one after another without the occurrence of returning behavior. Small values signal returning behavior to formerly visited pages such as overview and search result pages.

Centrality measures describe the importance of nodes in terms of how central they are located in the graph structure. For example, a particular node can be described by its degree centrality which can be interpreted as the amount to which users return to a specific page from varying other pages. Betweenness centrality for both nodes and edges measures whether there are bridging elements in the network structure, such as specific overview pages which a user frequently returns to in order to access other webpages. Therefore, both concepts can be seen as examples for comparison behavior of a user.

Eigenvector, katz and pagerank centrality indicate whether there is a wide choice of disjoint paths, where high average values indicate an interwoven structure of several important nodes. The intuition of closeness centrality and closeness vitality is that both measures assume high values if a node is located central in the whole network. For example, this applies to specific webpages a user has visited several times during the whole session. This extends the notion that there may be a specific number of pages that are central in the clickstream.

The summary statistics for the graph metrics of both shops is provided in the appendix.

Category	Metric	Feature	Description
Structure (8 metrics)	Number of nodes	NumberNodes*	Total number of nodes in the graph.
	Number of edges	NumberEdges*	Total number of edges in the graph.
	Number of cycle	NumberCircles	Total number of circles in the graph.
	Number of self-loops	SelfLoops	Total number of self-loops in the graph.
	Flow hierarchy	FlowHierarchy	Proportion of edges not being part of a cycle.
	Transitivity	Transitivity	The number of triangles in the graph divided by the maximum possible number of triangles.
	Density	Density	The sparseness in terms of connectivity for the whole graph.
	Average node connectivity	NodeConnectivity*	Average number of nodes for each distinct node pair that must be removed from the network in order to disconnect them.
Distance (6 metrics)	Average shortest path length	ShortestPath*	The average of the shortest path length for all distinct node pairs in the graph.
	Average eccentricity	Eccentricity*	Average of the longest shortest path for each single node in the graph.
	Diameter	Diameter*	The maximum eccentricity for the whole graph.
	Radius	Radius	The minimum eccentricity for the whole graph.
	Center	Center	Number of nodes with an eccentricity value equal to the radius.
	Periphery	Periphery	Number of nodes with an eccentricity value equal to the diameter.
Centrality (9 metrics)	Average in-degree / average out-degree	Degree*	Average of the number of edges converging from / to a node.
	Average neighbor degree	NeighborDegree	The average of the neighbor degree for each distinct node in the graph.
	Average closeness centrality	Closeness*	The average closeness, i.e. centrality of all nodes in the graph.
	Average closeness vitality	Vitality	The average change in closeness for all nodes if successively one node is removed from the graph.
	Average node betweenness centrality	NodeBetweenness	Importance of a node in terms of number of shortest paths passing through this node.
	Average edge betweenness centrality	EdgeBetweenness	Importance of an edge in terms of number of shortest paths passing through this edge.
	Average eigenvector centrality	Eigenvector	Different measures to compute the centrality of a node based on the adjacency matrix of the graph considering the linkage structure of the direct neighborhood of a node and partially a node's own edge structure.
	Average katz centrality	Katz*	
	Average pagerank centrality	Pagerank centrality*	

Table 2. Overview of our applied graph metrics (removed from the final feature set due to multicollinearity, see Chapter 4.2).*

3.3. Prediction Model Training and Assessment

We use prediction models to forecast whether a user session leads to a purchase. We set this target variable to one for all page views in a session if the user conducts a purchase during the session, and to zero otherwise. All predictive variables, i.e. the selected graph metrics and control variables introduced below, are normalized to their standard score to facilitate the interpretation of coefficients for the linear model in terms of their standard deviation from the mean.

We perform out-of-time validation and split our datasets sequentially into training and set; according to the month of the session. Data from September is used as training set whereas data from August is used as test set, resulting in an approximate split of 6:4 between training and test data. Out-of-sample in combination with out-of-time validation is commonly used in benchmarking studies to understand model performance in marketing (Linoff and Berry, 2011, p. 72) or credit scoring (Sobehart et al., 2000), where models are required to be stable over time. This is especially relevant in the e-commerce setting, since we want to test whether our model is able to predict the focal behavior for a different time period than the one in which the model was trained. The out-of-time validation approach is thus stricter in analyzing the performance of the model compared to randomized out-of-sample testing within the same period. We tune the meta-parameters of the prediction models introduced below by means of 5-fold cross validation on the training set. Since our data is highly imbalanced, we additionally applied synthetic minority over-sampling (SMOTE; Chawla et al., 2002), which creates artificial data points based upon the characteristics of a real observation of the minority class and its direct neighborhood to create a balanced dataset.

We select three different classification algorithms, a generalized linear logistic regression model (GLM) and two nonlinear tree-based models, which are random forest (RF) and gradient boosting machine (GB). We motivate the choice of logistic regression by its use in previous work (Byeon, 2013; Kalczynski et al., 2006). RF is chosen due its high performance in several forecasting benchmarks (e.g., Lessmann et al., 2015). We apply GB as a third method, because recent studies have found them to perform superior in similar classification tasks when compared to GLM and RF (Fitzpatrick and Mues, 2016). All models have the advantage that they are interpretable to a degree, which we use to examine the relative predictiveness of alternative graph metrics. The coefficients of logistic regression are interpretable in direction and size and allow significance testing. The RF and GB classifier provides variable importance scores, which also indicate the predictiveness of a variable (Breiman, 2001).

We provide two measures of prediction performance. We evaluate the models build on the respective variable sets using the area-under-the-precision-recall-curves (AUC-PR). The AUC-PR is commonly applied as a single-value metric similar to the area-under-the-ROC-curve (AUC; Fawcett, 2006) for model evaluation in case of imbalanced datasets (Saito and Rehmsmeier, 2015). The PR curve is constructed through pairwise plotting of precision and recall pairs at different classification thresholds, where recall is the proportion of observations predicted to be positive (i.e., purchase) in relation to all positive observations and precision is the rate of predictions that are correct. In general, the higher the value of AUC-PR, the better the model discriminates between the two classes. The second measure we apply is the lift index, which is a popular performance indicator for targeting models (Ling and Li, 1998). Under some assumption, lift is directly connected to the profitability of a targeting model (Martens and Provost, 2011; Piatetsky-Shapiro and Masand, 1999), which further motivates the choice of this performance indicator. The lift is based on a list of customer ordered according to their model-estimate conversion probability. In our case, the lift is defined as the share of hits, i.e. purchasers, in the top segment of $0<\theta<1$ of customers sorted by predicted purchase probability divided by the expected number of buyers in a random sample.

More formally, lift L_d is defined as:

$$L_d = \frac{\hat{\pi}_d}{\hat{\pi}} \quad (1)$$

with $\hat{\pi}_d$ denoting the fraction of purchasers among the top-d customers and $\hat{\pi}$ the prior probability of purchase, the lift assesses the degree to which a prediction model improves over a random benchmark.

To be able to assess the performance of our graph-based methodology in comparison to standard approaches, we will use an additional second feature set originating from the standard approach of feature extraction from clickstream (Table 3). Related to Table 1, we will use covariates of different categories such as *Page* and *Time*.

Feature	**Description**
SessionOverview	Number of visited pages of type 'overview' / 'product'/ 'sale' / 'search' in session.
SessionProduct	
SessionSale	
SessionSearch	
TabVisible	Is the tab currently visible?
Weekday	The weekday the session was started (1 – 7).
DayOfMonth	Day of the month (1 – 31) the session starts.
SessionStartHour	Hour of the session start (morning - midday - evening - night).
TimeOnPage	Time spent on page.
SessionTime	Total time of session.
PageVisitedBefore	Indicator whether the page has been visited before in the session.
Browser	The type of the browser the client uses.
ScreenSize	The screen size resolution of the visitor.
WindowSize	The window resolution of the visitor.
LocationZip	The zip code area of the city the user accesses the website from.
MajorCity	Indicator whether the website access happens from a major city.

Table 3. Overview of traditional features applied as a comparison approach to our graph approach.

4. Empirical Results

Based on the methodology discussed above, we report our empirical results in three steps. Firstly, we will take a detailed look at the correlation among the graph measures applied. Secondly, we analyze the performance of the tested classifiers based on AUC-PR and the lift measure. Finally, we will investigate the different graph measures in order to better understand their impact on the predictive accuracy.

4.1. Dataset Description

We use a two-month period of clickstream data of two large online retailers selling clothing and footwear, respectively. The data was collected from August to September 2015 and contains information such as identifiers (e.g., user id and session id), geographic- and user-based information (e.g., user agent) as well as path-, time- and behavioral-related information with regard to a customer’s journey on the respective website.

In the first step, we clean the data by deleting incomplete sessions and dismissing user sessions with less than four page views. Those sessions are referred to as bouncers which have no interest in the website in itself or generally to conduct a purchase. Furthermore, at least four clicks are necessary to complete the purchase process. With regard to potential bot elimination from the dataset we exclude one outlying user sessions with a length of 550 views, which we assume to be the product of automated website access.

The descriptive statistics of the final datasets are shown in Table 4. In total, the first shop contains 58,545 unique users performing a total of 692,975 page views. Of all 80,184 sessions, 4,256 sessions (approx. 5.31%) result in a purchase by a user. The second shop has a lower visitor count of 18,759 users who account for 32,850 sessions and 475,500 page views. Looking at the average value of page views for each visitor, users visiting the second shop on average look at more webpages per session compared to visitors of the first shop. Still, sessions of website visitors of the first shop result in more purchase conversions than in case of the second shop, where around 0.7 percent less purchases have taken place.

	Shop 1	**Shop 2**
Users	58,545	18,759
Sessions	80,184	32,850
Page views	692,975	475,500
Avg. page views	8.64	14.47
Purchase	4,256 (5.31%)	1,520 (4.63%)

Table 4. Descriptive overview of our final datasets.

4.2. Correlation Analysis of Graph Measures

In the first step, we calculate the correlation matrix of the graph metrics to understand which features embody similar information about the navigational structure of a user's journey on a website. From each set of highly collinear variables, we select only one variable for further analysis to avoid issues of multicollinearity. The corresponding correlation matrices for both datasets are shown in Figure 3. We see that within the three graph metrics categories – structural, distance and centrality – high correlation exists between subsets of the variables. Partly, this is not a surprising result since some measures are either variations of each other (e.g., eigenvector, katz and pagerank centrality) or their calculation is based upon another metric (e.g. eccentricity and the related metrics diameter and radius).

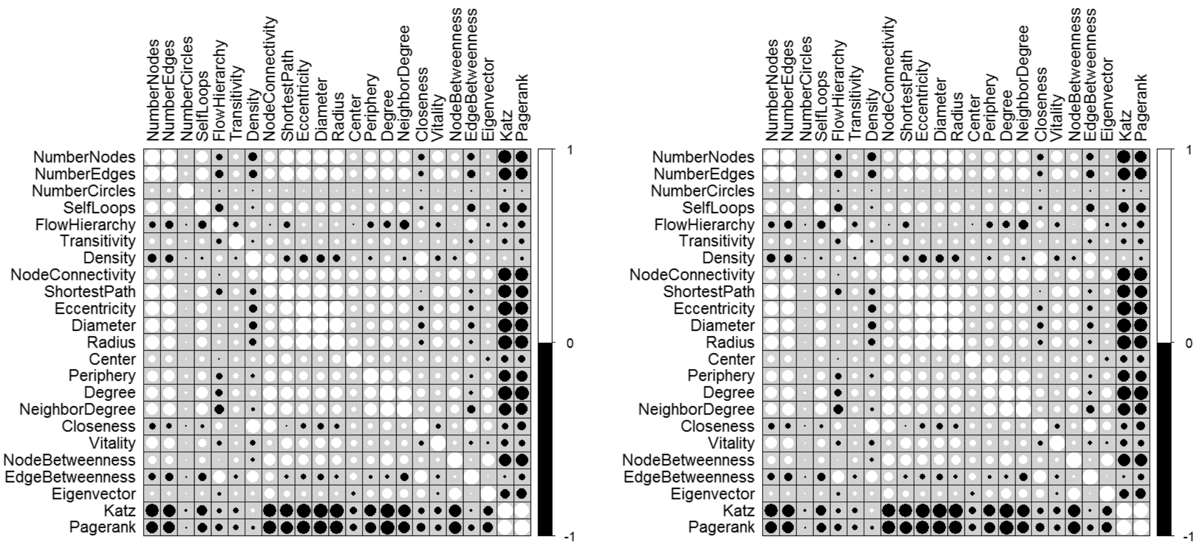

Figure 3. Correlation matrices for shop 1 (left) and shop 2 (right).

For both shops, we observe similar correlation patterns. The measure *NumberNodes* is highly correlated with *NumberEdges*. Additionally, the three metrics *Eccentricity*, *Radius* and *Diameter* contain almost the same informational content. Furthermore, the centrality measures katz and pagerank centrality – themselves highly correlated amongst each other - are negatively correlated with several graph metrics such as the structural components (number of nodes and edges) and the distance-based measures *ShortestPath*, *Eccentricity*, *Diameter* and *Radius*. The correlation of *NumberNodes* and *NumberEdges* can be interpreted in such a way that users tend to often perform as many click events as they visit unique webpages, i.e. webpages are generally visited only once and not several times by a user in a session. The correlation of *Eccentricity*, *Diameter* and *Radius* is unsurprising since they are based on the same basis, i.e. diameter being the maximum and radius being the minimum eccentricity in the graph.

To mitigate the issues of multicollinearity among the graph features, we remove highly correlated features on the basis of their variance inflation factor (VIF; Alin, 2010) calculated as

$$VIF_j = \frac{1}{1-R_j^2}\,, \tag{2}$$

where R^2 is the coefficient of determination from the regression of the covariate j on all other covariates (Stine, 1995). In contrast to the correlation coefficient, the VIF estimates the dependency of one covariate on all other covariates simultaneously, thus avoiding issues of the pairwise comparison. The higher the value of the VIF, the higher the correlation between the covariant j and all other variables. In general, covariates exceeding a VIF value between five and ten are seen as being prone to multicollinearity (Katrutsa and Strijov, 2017; Hair et al., 1998). We set our threshold to five and remove covariates exceeding this VIF value from the feature set. The calculation of the VIF is done in a step-wise manner, since the removal of a variable with high correlation affects the remaining variables influence. We recalculate the VIF for all remaining variables after removing the covariant with the highest VIF value from the preceding evaluation round. For both shops VIF results were almost consistent leading to the elimination of the covariates *Katz* (VIF = 2536.10 for shop 1 / 1201.07 for shop 2), *Diameter* (VIF = 176.74 / 758.24), *NumberNodes* (VIF = 166.33 / 200.10), *NodeConnectivity* (VIF = 82.06 / 45.88), *Closeness* (VIF = 30.37 / 22.65), *Pagerank* (VIF = 25.31 / 15.43), *ShortestPath* (VIF = 14.97 / 10.42), *Degree* (VIF = 10.59 / 5.08) and *NumberEdges* (VIF = 7.02 / 5.08) from the feature set. Additionally, in case of shop 1 the covariant *Eccentricity* (VIF = 993.34) and in case of shop 2 the *Radius* (VIF = 264.23) exceed the VIF threshold. Since these two metrics show high VIF values from the very beginning which drop significantly once either of the two is removed, we remove the covariant with the higher overall VIF, *Eccentricity,* (VIF = 993.34 for shop 1 / 257.25 for shop 2) for both shops and keep the covariant *Radius* in both datasets to increase consistency and facilitate the analysis.

This results in a final feature set consisting of 13 graph metrics, which we use for further analysis.

4.3. Predictive Performance

Using the subset of the 13 remaining graph features, we compare their predictive performance against the traditional feature set based on the GLM, RF and GB algorithms introduced above.

Looking at AUC-PR (Table 5), we observe that the graph-based approach outperforms the traditional set of variables in all six instances independent of the underlying model. We further observe that the RF performs worse compared to GB and the linear GLM for both shops. Furthermore, apart from the RF model, both models achieve higher AUC-PR values in case of shop 1. All models outperform the expected performance of a random model equal to the purchase rate of 5.3% and 4.6% for shop 1 and 2, respectively.

Model	**GLM**		**RF**		**GB**	
Covariates	**Graph**	**Traditional**	**Graph**	**Traditional**	**Graph**	**Traditional**
Shop 1	**0.372**	0.271	**0.287**	0.262	**0.372**	0.262
Shop 2	**0.311**	0.243	**0.300**	0.247	**0.317**	0.288

Table 5. AUC-PR values for shop 1 and shop 2 for the applied models.

For the lift measure, we observe that all three models trained on the applied graph metrics constitute for a clear improvement compared to random targeting. Figure 4 visualizes model lift in a gain chart for the three models on each dataset. Intuitively, the gain chart provides information about the number of purchasers if n% of users are targeted by the model, for example with a marketing incentive. Along the x-axis all views are plotted ordered by their predicted probability to purchase starting with those views having the highest probability. The y-axis represents the cumulative number of purchases among those page views. The upper grey bound shows the outcome of a perfect model which classifies all views according to their correct outcome, while a random model would result in a 45-degree diagonal. The steeper the curve is for a model, the better the model.

In case of the first shop, for the first around 30 percent of samples tested, both the GLM and the GB model perform almost equally, i.e. their performance in terms of classifying views with a high predicted purchase probability. In the beginning, RF performs slightly worse until around 30 percent of the samples are tested where the model exhibits a similar performance compared to the other two applied models but is soon visibly outperformed for larger samples. For the second shop, all three models are even more homogenous in terms of their predictive performance until a threshold of around 50 percent of samples is reached. Exceeding this threshold we observe again a similar performance of GLM and GB, while the RF model falls behind.

In general, the GLM model performs comparable in terms of lift compared to GB. This is surprising considering the general performance of the models and the ability of GB to model non-linear relations between the predictors. Given that all graph metrics are different measures to describe the same underlying graph structure, the good performance of the logistic regression model might be an indication that there are no significant non-linear dependencies between the graph metrics in predicting purchase behavior

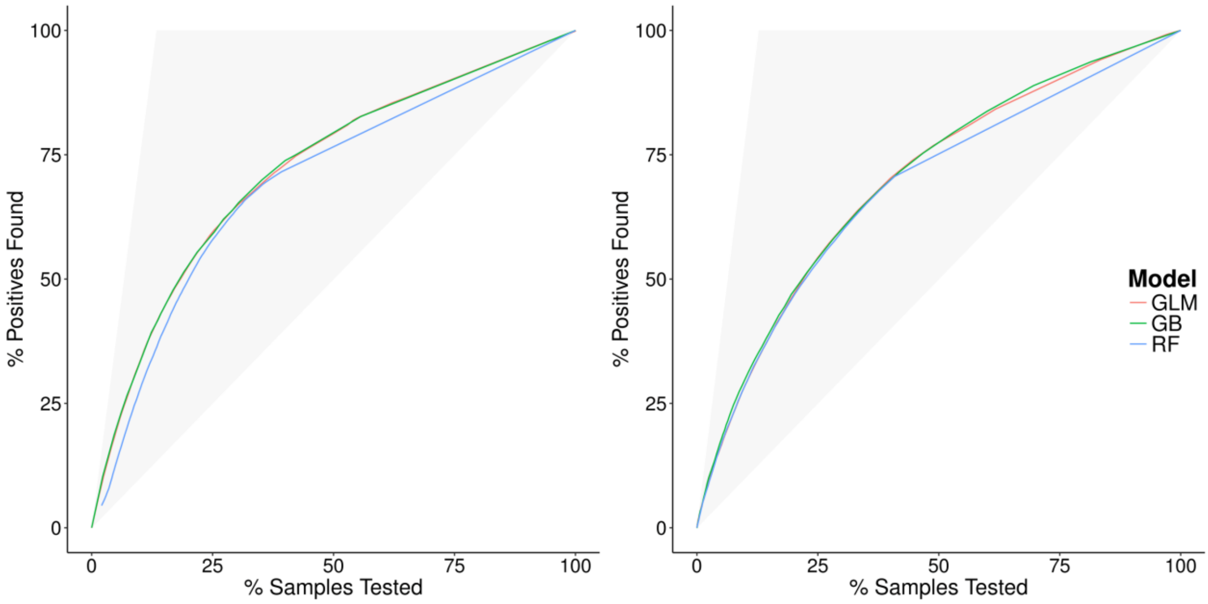

Figure 4. Lift chart for shop 1 (left) and shop 2 (right).

4.4. Variable Importance

In order to shed light on the direction of effect and performance of each graph metric, we analyze the model-wise importance of each graph measure. For the GLM model, we report the raw coefficients and odds ratios for each dataset (Table 6). Due to the large data size, we observe that almost all coefficients are highly significant even at the 0.01% level. The coefficient of the variable *NumberCircles* is least significant at the 0.1% level. We thus focus on the analysis of effect size.

Since all variables are standardized, we analyze their effect size expressed in terms of the impact a change by one standard deviation (SD) has on the odds ratio (Table 6). With the odds ratio defined by the ratio of probabilities $\mathrm{P}(Purchase)/\mathrm{P}(No\ purchase)$, an exponentiated effect above 1 indicates a larger purchase probability. For both shops, *Radius* and *SelfLoops* have a strong positive effect on purchase probability. Specifically, an increase in *Radius* by one standard deviation, associated with less compact graphs, indicates an increase in purchase odds by 135% (shop 1) or 51% (shop 2), while an increase in *SelfLoops* by one SD leads to an increase in purchase odds by 40% (shop 1) or 51% (shop 2). A slightly smaller effect exists for *EdgeBetweenness* where a one SD increase, due to less connections between nodes, is associated with a 12% (shop 1) or 28% (shop 2) increase. In contrast, we observe the largest negative impact on purchase odds for a decrease in *Density*, where a decrease by one SD, observed for sparser graphs, increases the odds of a purchase by 23% (1/0.81 = 1.23) (shop 1) or 39% (shop 2). *FlowHierarchy* is estimated to have a negative effect of similar size. While there are some differences in effect size, we observe no difference in direction for the above variables, which have the strongest impact. In sum, the observed pattern suggests that linear click-paths related to search behavior may be more indicative of users with purchase intention.

The variables *NumberCircles, Eigenvector NeighborDegree,* and *Center* show coefficients in different directions between shop 1 and 2, indicating that the underlying relationship may be shop dependent to a larger degree.

	GLM Model for Shop 1			**GLM Model for Shop 2**		
Variable	**Coefficient**	**Std. Error**	**Odds Ratio**	**Coefficient**	**Std. Error**	**Odds Ratio**
Intercept	-0.29 ***	0.004	0.75	-0.29 ***	0.005	0.75
NumberCircles	-0.04 **	0.014	0.96	0.09 ***	0.013	1.09
Density	-0.21 ***	0.006	0.81	-0.33 ***	0.009	0.72
Vitality	-0.12 ***	0.006	0.89	-0.06 ***	0.009	0.94
NodeBetweenness	0.05 ***	0.005	1.05	0.03 ***	0.006	1.03
EdgeBetweenness	0.11 ***	0.008	1.12	0.25 ***	0.010	1.28
Eigenvector	-0.06 ***	0.005	0.94	0.02 **	0.006	1.02
Radius	0.86 ***	0.007	2.35	0.41 ***	0.008	1.51
SelfLoops	0.34 ***	0.005	1.40	0.51 ***	0.007	1.66
FlowHierarchy	-0.20 ***	0.006	0.82	-0.25 ***	0.007	0.78
NeighborDegree	-0.15 ***	0.007	0.86	0.05 ***	0.006	1.06
Center	-0.10 ***	0.005	0.91	0.03 ***	0.006	1.03
Periphery	0.14 ***	0.005	1.15	0.05 ***	0.005	1.05
Transitivity	0.08 ***	0.004	1.09	0.11 ***	0.005	1.11
Significance levels: 0.0001 '***' 0.001 '**' 0.01 '*'						

Table 6. Estimated coefficients for the GLM model.

For the gradient boosted trees, we calculate the variable bag importance for both datasets based on the weighted increase in node purity for the splits on each variable averaged over all trees (Hastie et al., 2001). In other words, the variable importance captures the relative contribution to improve classification for each variable in the model. The variable importance scores reported in Figure 5 are scaled to sum up to 100 and are ordered according to their average importance for both shops.

The importance ranking for the non-linear GB model shows different patterns compared to the logit coefficient analysis in so far as *Density* and *FlowHierarchy* are only marginally relevant for purchase prediction while *Vitality*, *SelfLoops*, *NumberCircles* and *Radius* constitute the most important variables. Since high values of *Vitality* refer to the existence of important connections in the graph structure, the importance of the variable could be explained through being able to detect specific user behavior, i.e. signifying either goal-oriented, non-recursive browsing behavior or the existence of bridging elements in the user website journey such as overview or search result pages.

However, we observe that feature importance is centered around *Vitality* with a sharp decrease towards the second most important variable. Here we observe some deviation in variable importance between shops. In case of shop 1 the variables *Radius*, *Periphery* and *SelfLoops* are the next most important variables, whereas in case of shop 2 this is true for the variables *SelfLoops*, *NumberCircles* and *Center*. While the distance-based measures *Radius*, *Periphery* and *Center* refer to global characteristics of the clickstream graph, *SelfLoops* and *NumberCircles* are related to direct neighborhoods of single nodes. These two feature pairs could flag different browsing behaviors present in both shops.

All other variables only constitute for a small percentage in terms of variable importance and seem to be negligible for the distinction of purchasers and non-purchasers in case of the two datasets applied and the GB model.

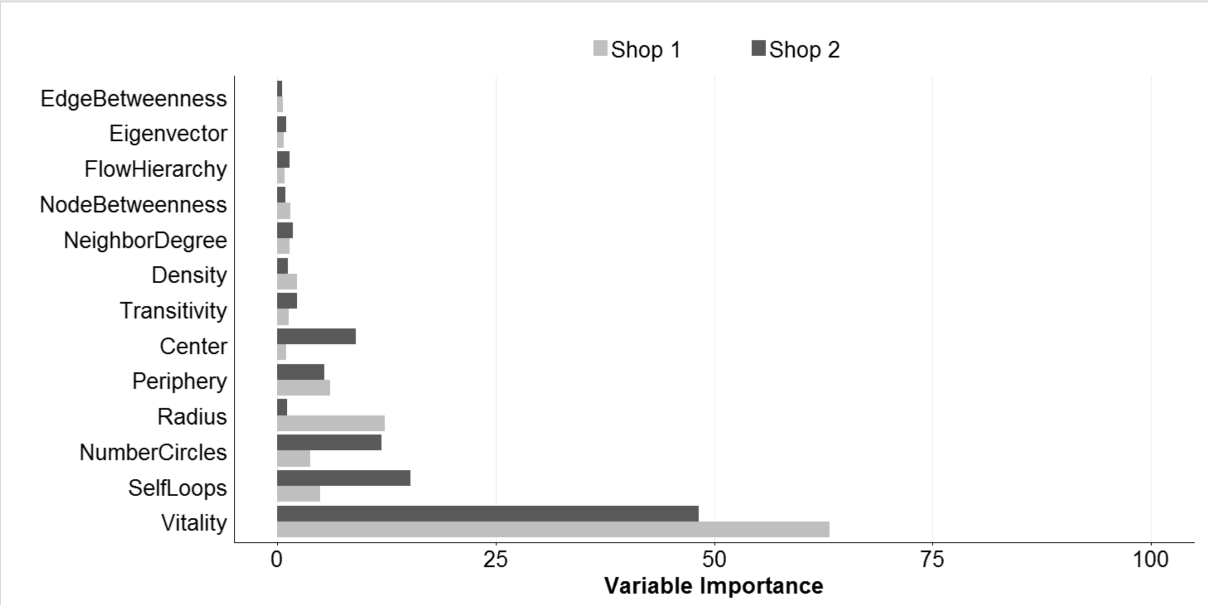

Figure 5. Variable importance for GB model for shop 1 and shop 2.

We use Partial Dependence Plots (PDP) for a deeper analysis of non-linear effects of each variable within the GB model (Hastie et al., 2001). PDP are a graphical tool to examine the marginal effect of each variable on the model prediction accounting for the (average) effects of all other variables. Figure 6 and Figure 7 show the PDP for shop 1 and shop 2, respectively. For both shops, we partly observe distinctive patterns, not completely in line with the general between-shops robustness of effects observed for GLM. Naturally, the PDP for both shops show the most distinctive patterns for the variables with the highest important scores (see Figure 5).

According to the PDP, an increasing value of *Vitality*, being the most important variable for both shops and which we interpret as a rising number of central pages in the user journey, shows similar behavior for both shops. Initially values below zero are linked to a high purchase probability but drop significantly once a value of zero is reached. However, after *Vitality* reaches a certain threshold the purchase probability increases again for both shops. Since this measure represents the change of distances for all present nodes in the graph, this metric might both be able to capture non-recursive type of browsing behavior and users with a high number of page views, both possibly being an indication of shopping behavior leading to a purchase.

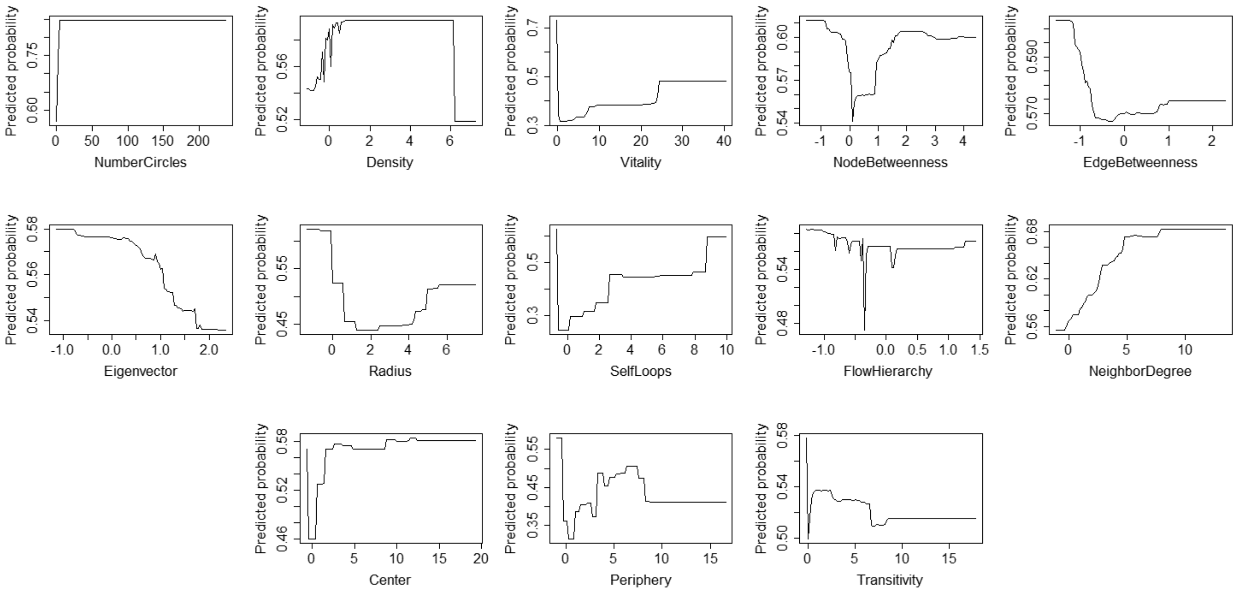

Figure 6. *Partial dependence plots for shop 1.*

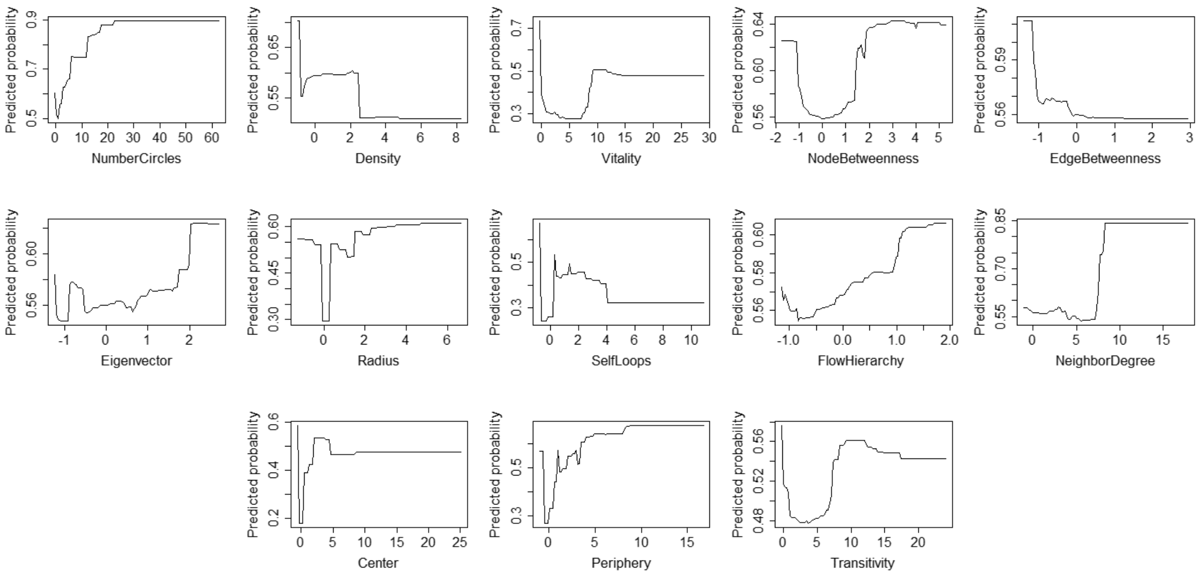

Figure 7. *Partial dependence plots for shop 2.*

The PDP of *SelfLoops*, which captures re-occurring webpage visits and constitutes the second most important variable, reveals a similar link. For both shops the purchase probability increases with the number of times a user revisits the same page. However, whereas in case of shop 1 the purchase probability continuously increases with a larger number of loops in the browser session this is not the case for shop 2. After reaching a value of around four the purchase probability decreases and remains stable from then on. Furthermore, an increasing value of *NumberCircles*, representing the existence of circles in the browsing structure of users, leads to an increasing purchase probability for both shops.

In case of shop 1, the two distance-based measures *Radius* and *Periphery* have been shown to be important for the prediction task. In general, the PDPs of both metrics resemble a U-shaped curve where both low and high values are associated with a higher purchase probability. For both metrics a drop occurs at a value of around zero where purchase probability resembles the lowest value.

In case of shop 2, the variable importance scores of *Center* and *Periphery* have been shown to be relevant for predicting purchasers as captured by the feature important scores. Despite a sharp drop at a value of zero, the PDP for *Center* illustrates the general relationship that the higher the value of this metric the more likely it is that a purchase occurs, signaling extensive browsing behavior within a direct neighborhood of a webpage. For *Transitivity* again a U-shaped curve is observable where the valley represents those users who generally do not conduct a purchase. Furthermore, rather low and values exceeding a value of five might signal two different shopping behaviors leading to a purchase which are goal-oriented and browsing-related shopping behavior.

Altogether these might be indicators that graph metrics are able to detect different shopping behaviors leading to a purchase which are either a goal-oriented or a browsing-related shopping experience. However, given the predictive value of the graph metrics, in-depth analysis beyond the scope of this paper will be necessary to identify the specific user intentions associated with a graph structure and could focus on experimental investigation of the link between stated user intention and each metric and establishing the robustness of the observed dependencies structures to different shops and product categories.

5. Conclusion

Using real-life clickstream datasets of two different shops we observe for both the linear GLM model and the non-linear Random Forest model that distance- and centrality-based graph metrics are effective in predicting purchase behavior of users. We derived user-centered, session-based graphs from clickstream data, where each graph is developed incrementally, i.e. each new page view of the user develops the graph further. Each of the 23 tested graph metrics are calculated for each intermediate state of a graph. We report and control for multicollinearity between the graph metrics by pre-processing using variable inflation factors and train three selected high-performing algorithms on the resulting dataset. Independent of the employed model, the proposed variables result in a substantial increase in the area-under-the-precision-recall-curve and model lift in predictive power compared to random targeting and a set of standard aggregation features derived from clickstream.

Looking at the importance of each graph metric, we observe clear differences in the relevance of variables between the linear and non-linear models. We suggest that closeness vitality in particular followed by radius and the number of self-loops and circles should be considered promising candidates in future applications.

We also identify some promising areas for future research. An alternative approach to calculate graph metrics could include different graph construction methods such as using bi-partite graphs, where two different types of nodes are included, to represent the structure of a user session in more detail. Additionally, constructing weighted graphs by rating frequently taken paths as more important or accounting for the time spent on specific pages could improve the representation of the users' journey on a website and consequently increase the accuracy when predicting the outcome of a session.

Appendix

Shop 1							
	Min.	**25% Quant.**	**Median**	**Mean**	**75% Quant.**	**Max.**	**Std. Dev.**
Purchase	0.00	-	-	0.14	-	1.00	-
NumberCircles	0.00	0.00	1.00	3.36	3.00	38150.00	132.95
Density	0.00	0.13	0.25	0.30	0.42	2.00	0.26
Vitality	0.00	1.00	7.50	54.51	36.00	25818.68	289.35
NodeBetweenness	0.00	0.00	0.15	0.12	0.17	0.50	0.09
EdgeBetweenness	0.00	0.13	0.20	0.21	0.28	0.50	0.14
Eigenvector	0.00	0.00	0.21	0.19	0.33	0.58	0.18
Radius	0.00	1.00	2.00	1.79	2.00	24.00	1.38
SelfLoops	0.00	0.00	0.00	0.67	1.00	13.00	1.02
FlowHierarchy	0.00	0.13	0.44	0.49	1.00	1.00	0.38
NeighborDegree	0.00	0.33	1.00	1.11	1.63	23.91	1.09
Center	1.00	1.00	1.00	1.58	2.00	20.00	0.85
Periphery	1.00	2.00	2.00	2.49	3.00	30.00	1.49
Transitivity	0.00	0.00	0.00	0.01	0.00	1.00	0.05

Table A. Summary statistics of graph metrics for shop 1.

Shop 2							
	Min.	**25% Quant.**	**Median**	**Mean**	**75% Quant.**	**Max.**	**Std. Dev.**
Purchase	0.00	-	-	0.13	-	1.00	-
NumberCircles	0.00	1.00	2.00	7.49	7.00	14298.00	117.34
Density	0.00	0.09	0.18	0.25	0.33	2.00	0.23
Vitality	0.00	4.00	24.29	225.33	125.17	29824.62	897.00
NodeBetweenness	0.00	0.09	0.14	0.12	0.17	0.50	0.08
EdgeBetweenness	0.00	0.09	0.15	0.18	0.22	0.50	0.12
Eigenvector	0.00	0.00	0.18	0.19	0.32	0.58	0.16
Radius	0.00	1.00	2.00	2.76	4.00	28.00	2.32
SelfLoops	0.00	0.00	1.00	1.09	1.00	37.00	1.64
FlowHierarchy	0.00	0.12	0.33	0.42	0.67	1.00	0.35
NeighborDegree	0.00	0.80	1.39	1.79	2.23	48.61	2.03
Center	1.00	1.00	1.00	1.62	2.00	31.00	1.03
Periphery	1.00	2.00	2.00	2.89	3.00	39.00	1.98
Transitivity	0.00	0.00	0.00	0.01	0.00	1.00	0.04

Table B. Summary statistics of graph metrics for shop 2.

References

Alin, A. (2010). Multicollinearity. Wiley Interdisciplinary Reviews: *Computational Statistics*, 2(3), 370-374.

Anitha, A. (2010). A new web usage mining approach for next page access prediction. *International Journal of Computer Applications*, 8(11), 7-10. doi:10.5120/1252-1700

Antonellis, P., Makris, C., and Tsirakis, N. (2009). Algorithms for clustering clickstream data. *Information Processing Letters*, 109(8), 381-385. doi:10.1016/j.ipl.2008.12.011

Banerjee, A., and Ghosh, J. (2001). Clickstream clustering using weighted longest common subsequences. *Proceedings of the Web Mining Workshop at the 1st SIAM Conference on Data Mining* (pp. 33-40).

Berka, P., and Labský, M. (2007). Predicting page occurrence in a click-stream data: statistical and rule-based approach. *Advances in Data Mining. Theoretical Aspects and Applications* (pp. 135-147). Springer Berlin Heidelberg.

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5-32.

Byeon, H. (2013). Evaluating the online buying behavior using network analysis. *International Journal of Advancements in Computing Technology*, 5(12).

Chan, T., Joseph, I., Macasaet, C., Kang, D., Hardy, R. M., Ruiz, C., Porras, R., Baron, B., Qazi, K., Hannon, P. and Honda, T. (2014). Predictive models for determining if and when to display online lead forms. *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (pp. 2882-2889). AAAI Press.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). *SMOTE: synthetic minority over-sampling technique*. Journal of artificial intelligence research, 16, 321-357.

Colladon, A. F., and Remondi, E. (2017). Using social network analysis to prevent money laundering. *Expert Systems with Applications*, 67, 49-58. doi: 10.1016/j.eswa.2016.09.029

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861-874. doi:10.1016/j.patrec.2005.10.010

Fitzpatrick, T., and Mues, C. (2016). An empirical comparison of classification algorithms for mortgage default prediction: evidence from a distressed mortgage market. *European Journal of Operational Research*, 249(2), 427-439. doi:10.1016/j.ejor.2015.09.014

Girija, P., and Kavitha, V. (2013). An Approach for predicting user's web access pattern. *International Journal of Computer Science and Management Research*, 2(5), 2585-2589.

Gündüz, Ş., and Özsu, M. T. (2003). A web page prediction model based on click-stream tree representation of user behavior. *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 535-540). ACM.

Hagberg, A. A., Schult, D. A. and Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX", *Proceedings of the 7th Python in Science Conference* (pp. 11–15*)*, Gäel Varoquaux, G., Vaught, T. and Millman, J. (Eds), Pasadena, CA USA.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., and Tatham, R. L. (1998). *Multivariate data analysis,* 5(3), Upper Saddle River, NJ: Prentice hall, 207-219.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning* (2nd ed.). New York: Springer series in statistics.

He, Y. L., Liu, J. N., Hu, Y. X., and Wang, X. Z. (2015). OWA operator based link prediction ensemble for social network. *Expert Systems with Applications*, 42(1), 21-50. doi: 10.1016/j.eswa.2014.07.018

Hong, T., and Kim, E. (2012). Segmenting customers in online stores based on factors that affect the customer's intention to purchase. *Expert Systems with Applications*, 39(2), 2127-2131. doi: 10.1016/j.eswa.2011.07.114

Iwanaga, J., Nishimura, N., Sukegawa, N., and Takano, Y. (2016). Estimating product-choice probabilities from recency and frequency of page views. *Knowledge-Based Systems*, 99, 157-167. doi:10.1016/j.knosys.2016.02.006

Jiang, Q., Tan, C. H., and Wie, K. K. (2012). Cross-website navigation behavior and purchase commitment: A pluralistic field research. *Proceedings of the 16th Pacific Asia Conference on Information Systems (PACIS)*.

Kalczynski, P. J., Senecal, S., and Nantel, J. (2006). Predicting on-line task completion with clickstream complexity measures: A graph-based approach. *International Journal of Electronic Commerce*, 10(3), 121-141. doi:10.2753/jec1086-4415100305

Katrutsa, A., and Strijov, V. (2017). Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. *Expert Systems with Applications*, 76, 1-11.

Kitts, B., Hetherington-Young, K., and Vrieze, M. (2002). Large-scale mining, discovery and visualization of WWW user clickpaths. *International Journal of Image and Graphics*, 02(01), 21-48. doi:10.1142/s0219467802000536

Lee, H., Choi, S. Y., and Kang, Y. S. (2009). Formation of e-satisfaction and repurchase intention: Moderating roles of computer self-efficacy and computer anxiety. *Expert Systems with Applications*, 36(4), 7848-7859. doi:10.1016/j.eswa.2008.11.005

Lee, M., Ferguson, M. E., Garrow, L. A., and Post, D. (2010). The impact of leisure travelers' online search and purchase behaviors on promotion effectiveness. Working Paper, Georgia Institute of Technology.

Lessmann, S., Baesens, B., Seow, H., and Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124-136. doi:10.1016/j.ejor.2015.05.030

Ling, C. X., and Li, C. (1998). Data mining for direct marketing: Problems and solutions. In R. Agrawal, P. E. Stolorz, G. Piatetsky-Shapiro (Eds.). *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining* (pp. 73–79). AAAI Press, Menlo Park.

Linoff , G. S. and Berry, M. J. A. (2011). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management* (3rd ed.). Wiley Publishing.

Lu, L., Dunham, M., and Meng, Y. (2005). Mining significant usage patterns from clickstream data. *Advances in Web Mining and Web Usage Analysis* (pp. 1-17). Springer Berlin Heidelberg.

Martens, D., and Provost, F. (2011). Pseudo-social network targeting from consumer transaction data. Faculty of Applied Economics, University of Antwerp.

Moe, W. W. (2003). Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *Journal of Consumer Psychology*, 13(1-2), 29-39. doi:10.1207/s15327663jcp13-1&2_03

Moe, W. W., and Fader, P. S. (2004). Capturing evolving visit behavior in clickstream data. *Journal of Interactive Marketing*, 18(1), 5-19. doi:10.1002/dir.10074

Moe, W. W., Chipman, H., George, E. I., and McCulloch, R. E. (2002). A Bayesian treed model of online purchasing behavior using in-store navigational clickstream. Working paper, University of Texas at Austin, Austin.

Montgomery, A. L., Li, S., Srinivasan, K., and Liechty, J. C. (2004). Modeling online browsing and path analysis using clickstream data. *Marketing Science*, 23(4), 579-595. doi: 10.1287/mksc.1040.0073

Óskarsdóttir, M., Bravo, C., Verbeke, W., Sarraute, C., Baesens, B., and Vanthienen, J. (2017). Social network analytics for churn prediction in telco: Model building, evaluation and network architecture. To appear in: *Expert Systems with Applications*. doi: 10.1016/j.eswa.2017.05.028

Padmanabhan, B., Zheng, Z., and Kimbrough, S. O. (2006). An empirical analysis of the value of complete information for eCRM models. *MIS Quarterly*, 247-267.

Pai, D., Sharang, A., Yadagiri, M. M., and Agrawal, S. (2014). Modelling visit similarity using click-stream data: A supervised approach. *Web Information Systems Engineering – WISE* (pp. 135-145). Springer International Publishing.

Panagiotelis, A., Smith, M. S., and Danaher, P. J. (2014). From Amazon to Apple: Modeling online retail sales, purchase incidence, and visit behavior. *Journal of Business & Economic Statistics*, 32(1), 14-29. doi: 10.1080/07350015.2013.835729

Park, C. H., and Park, Y. (2016). Investigating purchase conversion by uncovering online visit patterns. *Marketing Science*, 35(6), 894-914. doi:10.1287/mksc.2016.0990

Park, S., Suresh, N. C., and Jeong, B. K. (2008). Sequence-based clustering for Web usage mining: A new experimental framework and ANN-enhanced K-means algorithm. *Data & Knowledge Engineering*, 65(3), 512-543. doi: 10.1016/j.datak.2008.01.002

Piatetsky-Shapiro, G., and Masan, B. (1999). Estimating campaign benefits and modeling lift. *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining*. San Diego, CA, USA.

Pitman, A., and Zanker, M. (2010). Insights from applying sequential pattern mining to e-commerce click stream data." In: *IEEE International Conference on Data Mining Workshops (ICDMW)* (pp. 967-975). IEEE.

Saito, T., and Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS One*, 10(3), e011843.

Sarwar, S. M., Hasan, M., and Ignatov, D. I. (2015). Two-stage cascaded classifier for purchase prediction." arXiv preprint arXiv:1508.03856.

Sato, S., and Asahi, Y. (2012). A daily-level purchasing model at an e-commerce site. *International Journal of Electrical and Computer Engineering*, 2(6). doi:10.11591/ijece.v2i6.1816

Senecal, S., Kalczynski, P. J., and Fredette, M. (2014). Dynamic identification of anonymous consumers visit goals using clickstream. *International Journal of Electronic Business*, 11(3), 220. doi:10.1504/ijeb.2014.063036

Shams, B., and Haratizadeh, S. (2017). Graph-based collaborative ranking. *Expert Systems with Applications*, 67, 59-70. doi:10.1016/j.eswa.2016.09.013

Sismeiro, C., and Bucklin, R. E. (2004). Modeling purchase behavior at an e-commerce web site: A task-completion approach. *Journal of Marketing Research*, 41(3), 306-323. doi:10.1509/jmkr.41.3.306.35985

Stange, M., and Funk, B. (2015). How much tracking is necessary? – The learning curve in bayesian user journey analysis. *Proceedings of the 23rd European Conference on Information Systems*. Muenster, Germany.

Statista (2016a). Global retail e-commerce sales 2014-2020 | Statistic. Retrieved May 22, 2017, from http://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/
Statista (2016b). Global e-retail growth rate 2020 | Statistic. Retrieved May 22, 2017, from http://www.statista.com/statistics/288487/forecast-of-global-b2c-e-commerce-growt/
Stine, R. A. (1995). Graphical interpretation of variance inflation factors. *The American Statistician*, 49(1), 53-57.
Suh, E., Lim, S., Hwang, H., and Kim, S. (2004). A prediction model for the purchase probability of anonymous customers to support real time web marketing: a case study. *Expert Systems with Applications*, 27(2), 245-255. doi:10.1016/j.eswa.2004.01.008
Van den Poel, D., and Buckinx, W. (2005). Predicting online-purchasing behaviour. *European Journal of Operational Research*, 166(2), 557-575. doi:10.1016/j.ejor.2004.04.022
Van der Meer, D., Dutta, K., Datta, A., Ramamritham, K., and Navanthe, S. B. (2000). Enabling scalable online personalization on the web. *Proceedings of the 2nd ACM Conference on Electronic Commerce* (pp. 185-196). ACM.
Vroomen, B., Donkers, B., Verhoef, P. C., and Franses, P. H. (2005). Selecting profitable customers for complex services on the Internet. *Journal of Service Research*, 8(1), 37-47. doi:10.1177/1094670505276681
Wu, F., Chiu, I. H., and Lin, J. R. (2005). Prediction of the intention of purchase of the user surfing on the web using hidden markov model. In: *Proceedings of ICSSSM'05 International Conference on Services Systems and Services Management.* IEEE, pp. 387-390.
Zhang, Y., Bradlow, E. T., and Small, D. S. (2015). Predicting customer value using clumpiness: From RFM to RFMC. *Marketing Science*, 34(2), 195-208. doi:10.1287/mksc.2014.0873
Zhao, Y., Yao, L., and Zhang, Y. (2016). Purchase prediction using Tmall-specific features. *Concurrency and Computation: Practice and Experience*, 28(14), 3879-3894. doi:10.1002/cpe.3720
Zheng, Z., Padmanabhan, B., and Kimbrough, S. O. (2003). On the existence and significance of data preprocessing biases in web-usage mining. *INFORMS Journal on Computing*, 15(2), 148-170. doi:10.1287/ijoc.15.2.148.14449

ARTICLE 9:

THE PRICE OF PRIVACY: AN EVALUATION OF THE ECONOMIC VALUE OF COLLECTING CLICKSTREAM DATA[11]

Authors: Annika Baumann, Humboldt-Universität zu Berlin, Berlin, Germany

Johannes Haupt, Humboldt-Universität zu Berlin, Berlin, Germany

Fabian Gebert, Akanoo GmbH, Hamburg, Germany

Stefan Lessmann, Humboldt-Universität zu Berlin, Berlin, Germany

Submitted to: Business & Information Systems Engineering (BISE), Invitation to Resubmit.

Abstract

The analysis of clickstream data facilitates the understanding and prediction of customer behavior in e-commerce. Companies can leverage such data to increase revenue. For customers and website users, on the other hand, the collection of behavioral data entails privacy invasion. The objective of the paper is to shed light on the trade-off between privacy and the business value of customer information. To that end, we review approaches to convert clickstream data into behavioral traits, which we call clickstream features, and propose a categorization of these features according to the potential threat they pose to user privacy. We then examine the extent to which different categories of clickstream features facilitate predictions of online user shopping patterns and approximate the marginal utility of using more privacy adverse information in behavioral prediction models. This way, the paper links the literature on user privacy to that on e-commerce analytics and takes a step toward an economic analysis of privacy costs and benefits. In particular, the results of empirical experimentation with large real-world e-commerce data suggest that the inclusion of short-term customer behavior based on session-related information leads to large gains in predictive accuracy and business performance, while storing and aggregating usage behavior over longer horizons has comparably less value.

Keywords: *Predictive analytics, E-Commerce, Privacy, Behavioral Targeting, Clickstream Data*

[11] This article is provided with kind permission from Springer Nature. The original version is available at: https://doi.org/10.1007/s12599-018-0528-2.

1. Introduction

The e-commerce sector is steadily growing and estimated to have reached $1.915 trillion of sales turnover worldwide in 2016 (eMarketer 2016). With customer increasing spending, web usage mining has been established as a common practice by e-shops to offer website visitors an enhanced user experience and to better understand customer behavior (Cooley et al. 1997). The underlying data are collected in the form of clickstreams, which might include information such as the pages visited and the time spent on each page (Senécal et al. 2005). Clickstream data is seen as one of the top value adding data sources by businesses (Statista 2016a) with applications in online marketing, customer analysis, or website development. Within online marketing, clickstream mining has been readily adopted by business and academia to understand the behavior of website visitors. Use cases of individual-level clickstream data include customer targeting (e.g., Pai et al. 2014), understanding navigational preferences (e.g., Montgomery et al. 2004), and predicting customer conversion (e.g., Buckinx and Van den Poel 2005). But since no good comes without harm, the collection of user data always brings with it the possible hazard of privacy related issues, which pose ethical and economic risks to both customers and companies.

The informational privacy of website visitors is of concern for e-shops because the success of converting the visitors into customers (or lack thereof) depends, amongst other things, on the potential risks of the transaction as perceived by the visitor (Metzger 2004). From a user perspective, perceived risks of privacy exist in the form of third-party access to personal information, misuse of exposed information, unconsented secondary use of provided information and unintended mining or mapping of individual behavior (Dinev et al. 2013). The perceived risk of e-commerce transactions can be mitigated and user decisions positively influenced by increasing trust in the website through comprehensive privacy protection (Kim et al. 2008, Nofer et al. 2014).

One way to improve perceived privacy is to avoid use of user data unless it has been provided willingly by the customer (Liu et al. 2005). Clickstream data on the other hand is collected without action or consent by the website visitor. Consequently, privacy concerns do not only include existing customers who need to provide their sensitive personal information to complete the buying process, but also prospective customers who are anonymous and have not actively provided consent for the use of their data. In addition, online advertising companies such as DoubleClick collect user data in form of clickstream across the users' whole browsing history, combining several data sources and therefore intervening with their privacy in order to offer them the most fitting advertisements based on their aggregated website visits and search engine requests (Akrivopoulou and Stylianou 2009, p.125).

In general, privacy preserving data collection and analysis has been in the center of attention of big data research (e.g., Agrawal and Srikant 2000). However, as we detail in Chapter 3, there is still a lack of research focusing on the collection of clickstream data and the prediction of customer behavior under the restriction to simultaneously maintain a certain level of privacy. We argue that the collection of customer data is a strategic business decision and needs to be evaluated according to its marginal gain in relation to incurred risks and costs by managers and customers alike. Since the amount and type of data collected and stored is in the control of the e-shop and clickstream data is dispensable for the direct operational sales processes, the strategic question is what level of privacy in data collection is suitable to maximize sales performance under minimum risk exposure.

To answer this question, we review approaches to convert raw clickstream data into behavioral traits, which we call clickstream features, and identify groups of clickstream features based on their relevance for privacy issues. We then examine the economic value of clickstream features from different privacy categories through the lens of predictive modeling. In particular, we consider an e-commerce context and assume a company to gather clickstream data with the intention to predict customer behavior. Accurate behavior predictions can, for example, inform the company's marketing strategy and, more generally, aid in achieving growth targets. Drawing upon the literature on cost-sensitive learning, we link the economic value of clickstream data to the accuracy of a behavior prediction model. This allows us to quantify the marginal profit gain associated with employing a set of clickstream features and the opportunity costs of refraining from using these features, respectively.

So far, existing research considering the privacy aspect of clickstream data collection focused on whether several data sources (Padmanabhan et al. 2006) or a larger amount of data comprising a longer observation period (Stange and Funk 2015) yield advantages in predictive accuracy. We contribute to existing literature by focusing on what kind of clickstream features need to be included in a predictive model to obtain sufficiently accurate conversion predictions based on empirical evidence for two e-shops. Additionally, we provide an economic analysis of the privacy-accuracy trade-off to inform managerial decision-making. For example, we show that the inclusion of short-term clickstream data derived from session-related information leads to large gains in targeting accuracy, while long-term-based clickstream features over several sessions facilitates only a marginal gain in accuracy and value for the observed shops.

The remainder of the paper is organized as follows: Chapter 2 discusses the background and motivation of our work. Chapter 3 reviews related literature. Chapter 4 explains our methodology. Empirical results are presented and discussed in Chapter 5. Chapter 6 concludes the paper, states limitations and gives an outlook to future research.

2. Background and Motivation

In this section, we will discuss the concept of clickstream data collection in more detail and highlight its relevance for privacy-related aspects.

In general, the collection of user data on the Internet occurs in two distinct ways. Internet users may provide information actively and consciously, e.g. by creating a user account or by conducting a transaction where process completion requires the provision of personal information. They also pass information passively as a byproduct of visiting webpages in that every visit – or 'click' – leaves a digital footprint that is stored in web server logfiles and, in conjunction with subsequent page visits, provides what is called clickstream data (Skok 2000). Clickstream can be defined as a "record [which contains information about] the Internet service provider, the type of computer and software used, the website linked from, the amount of time spent perusing each page, and exactly what parts of the website were explored and for how long" (Solove 2001). Therefore, clickstream might not only include information about the path a user has taken through the website but also details about the interaction with the website in form of click-, scroll-, tab switch and basket events. Additionally, user agent data transferred with the clickstream such as the access device, browser information and screen resolution can be derived from it.

Figure 1 shows an example of three sessions of users visiting a website and how clickstream is collected in this process. A session is comprised of a series of webpages visited by a user that is terminated when no interaction takes place for a specified duration. In this example, each user takes a different path through the website. At each single traversal through a webpage data is gathered the form of clickstream. For example, when user A first visits the website the overall visit count is set to one and the number of webpages visited is updated at each page traversal, i.e. the webpage count is set to one when visiting the first webpage, set to two when the second webpage is visited and so on until the user leaves the website.

Informational bits can also be continued based upon historical clickstream data. For example, once user A re-visits the website the visit count is then updated to two.

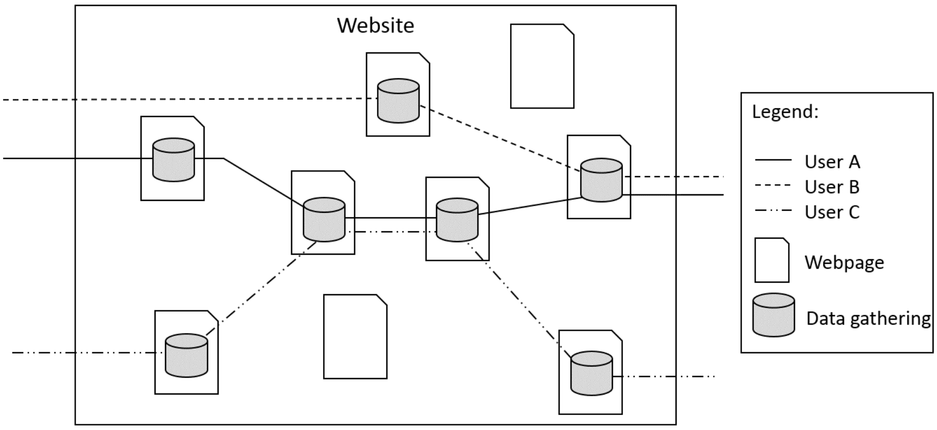

Figure 1. *Three examples of the clickstream data collection process.*

Personal data is protected under the aspect of informational privacy, which is defined as "the individual interest in avoiding disclosure of personal matters" (Lin 2002, p. 1094). Informational privacy has become especially relevant in the new area of the Internet and information technology where the collection and processing of data became beyond measure. In general, the collection and use of Internet user data is regulated in different ways depending on the country. For example, regulation in the US is sparse, while the European Union requires websites to obtain the user's permission regarding cookie placement and informing them whether data is collected and how it is used (Baumer et al. 2004). Legal restrictions define data security standards for certain types of data according to the sensitivity of the information, e.g. anonymous, personally identifiable information (PII), or medical data[12]. Data is considered as personally identifiable when a connection between the data and an individual is possible with reasonable effort. Such PII might be for example an e-mail address, a name, telephone number or other identifiers such as a social security number (Lin 2002). Clickstream is not classified as PII but still poses privacy threats such as potential de-anonymization, secondary use of data, unknown extent of data collection and the possibility to combine non-PII clickstream data with personal data (Sipior et al. 2011; Pollach 2011).

In this regard, many Internet users are not aware of the information they transmit while browsing and what kind of data is collected by whom (Hoofnagle et al. 2012). They are left with the feeling as if they "lost all control over how personal information is collected and used by companies" (Turow et al. 2009). Users who are not registered or logged into a site can be considered as anonymous by choice. Nevertheless, their clickstream data is collected and used to track their behavior when visiting a website. From a shop owners' perspective, motivations to do so include developing user profiles, for example to inform marketing actions. Given that anonymous visitors have not agreed to the collection and use of their clickstream data, they may hold a certain "expectation of privacy in clickstream data" (Skok 2000).

Clickstream data can constitute a severe threat with respect to website visitors' privacy. For example, clickstream data has been shown to facilitate the de-anonymization and access to personal information of users through revealing URLs (Libert 2015; Greis 2016). Since clickstream data contains the URLs of the webpages a user has visited, it is possible to track what is of interest for a specific user. Here, strongly sensitive information such as personal preferences or healthcare information can be revealed

[12] For example, see the *Health Insurance Portability and Accountability Act of 1996* or the California *Online Privacy Protection Act of 2003* for the US or the *General Data Protection Regulation* for EU regulation.

when this specific information is part of a URL (Libert 2015). Furthermore, URLs which contain account access information such as an e-mail address, being classified as PII, could be also revealed through non-hidden URLs (Greis 2016). In the e-commerce setting the URLs of e-shops might reveal in what kind of sensible products a user might be interested in and could possibly combined with personal information among log-in.

In addition, the long-term observation of behavioral user patterns can be used to de-anonymize users by matching recurring visit and page interaction patterns, collected in the past or on other websites, to an anonymous visitor (Yang 2010). Here, behavioral patterns such as the specific journey the user takes on the website, how long she stays on specific pages and where click and scroll events take place might be an indication who is visiting the website through the match of reoccurring patterns. In this case a user can even be de-anonymized when cookie deletion takes place since no identifier in the form of an ID is necessary. However, this approach is only applicable when a lot of data is available (Yang 2010). Another method, which can constitute a threat to online user privacy is browser fingerprinting. Research shows a high success rate of browser (re-)identification on the basis of user agent information (e.g., Eckersley 2010; Nikiforakis et al. 2014). Here, specific information about which browser a website visitor uses in combination with the underlying version is often so unique that single users can be identified based upon the user agent information collected altogether with the clickstream data. Since no long-term observation and no revealing URLs are necessary, this can be seen as the most obtrusive approach.

These cases illustrate how the collection of clickstream data may impede user privacy. More specifically, they show how raw clickstream data can be converted into features that characterize and potentially predict user behavior, which can be considered an invasion of user privacy in itself. In combination with increasing privacy awareness by consumers, data privacy statements have also become a part of trust-related marketing communications for companies (Bansal et al. 2015). Consequently, management has an incentive to reflect the degree to which they collect and store sensitive customer data.

3. Related Research

Using clickstream data as a means to predict a specific object of interest has been widely adopted in the literature. Possible prediction targets include the likelihood of customer churn (Moertini and Ibrahim 2015), user personalization approaches (e.g., Pai et al. 2014), or the prediction of purchase behavior and conversion (e.g., Buckinx and Van den Poel 2005). We provide an overview on relevant literature in the field of conversion prediction from two perspectives which are the features used for prediction and in what regard the privacy aspect in relation with clickstream data has been considered by literature so far.

3.1. Conversion Prediction and Clickstream Features

This section will give a detailed overview of features extracted from raw clickstreams to predict conversion as a basis for our own set of clickstream in Section 4.1. We focus on previous work related to conversion modeling because purchase prediction is one of the most common fields in prior literature. Furthermore, since conversion (e.g., a purchase) occurs on a single website, clickstream data collection and privacy are under direct control of the site owner; as opposed to online advertisement, where data collection routinely involves third party providers such as ad networks (e.g., Stange and Funk 2014).

Table 1 provides on overview on related literature focusing on the features used for predictive modelling. We group those features into classes depending on whether they belong to clickstream data or additional information. Clickstream features are further sub-grouped into the more fine-grained categories Page, Time, Monetary, Page Interaction and User Agent. Furthermore, we highlight those papers which have a focus on one of the main topics of our paper which is whether they cover a privacy and/or a profit analysis.

Existing research in predictive modelling made use of a number of clickstream features which we group into five categories. The first three, *Page*, *Time* and *Monetary*, are based on the well-known concept of recency, frequency and monetary value analysis (Zhang et al. 2015). *Page* combines data related to the path a website visitor traverses and how often specific pages or page categories have been visited. *Time* contains information about the time spent on each page or aggregated page categories. *Monetary* collects outcomes of historical and present purchase behavior. The monetary value of the purchase can be taken from clickstream data as a form of basket value. Aggregated amounts of basket values in the clickstream data sum up to historical purchase information of website visitors. The fourth category, which we call *page interaction,* includes variables related to basket actions (e.g., an item is placed in the basket during a session of a user), click, scroll and tab switch events. The last clickstream feature category, *user agent*, consists of information related to the access device, browser, screen resolution and IP-resolved location. The additional category *demographics*, which does not belong to clickstream data, but is included in our survey to derive a comprehensive picture of the feature categories used in literature. Demographics contains, for example, data related to gender, income and education of a user. We use these five categories to classify prior work in conversion modeling in terms of the employed data (Table 1).

Furthermore, with respect to the temporal reference of the clickstream features, we note that varying time horizons of clickstream features have been in the focus of research, which is also depicted in Table 1. The data used can be solely based on the current user session or alternatively can contain information across sessions, capturing historical information in terms of earlier website visits and purchases.

Reference	Privacy Focus	Profit / Business Value	Feature Horizon		Feature Category					
					Clickstream					Additional information
			Current Session	Across Session	Page	Time	Monetary	Page Interaction	User Agent	Demographics
This Paper	x	x		x	x	x	x	x	x	
Banerjee and Ghosh 2001			x		x	x				
Chan et al. 2014				x	x			x		x
Iwanaga et al. 2016				x	x	x				
Jiang et al. 2012			x		x	x				
Lee et al. 2010		x		x	x	x				
Lu et al. 2005			x		x					
Moe 2003			x		x	x				
Moe and Fader 2004				x		x	x			
Moe et al. 2002		Lift	x		x	x				
Padmanabhan et al. 2006	x	Lift		x	x	x	x			x
Park and Park 2015			x		x					
Pitman and Zanker 2010			x		x			x		
Sarwar et al. 2015			x		x	x	x			
Sato and Asahi 2012				x	x		x			
Senécal et al. 2014			x		x	x		x		
Sismeiro and Bucklin 2004		x		x	x	x		x		
Stange and Funk 2015	x			x	x	x	x	x		
Suh et al. 2004		Lift	x		x	x				
Buckinx and Van den Poel 2005				x	x	x	x			x
Vroomen et al. 2005		x		x		x	x	x		x
Wu et al. 2005			x		x					
Zhao et al. 2016				x	x	x	x			
Zheng et al. 2003		Lift		x	x	x	x			x

Table 1. Overview of focus, feature categories and time horizons used in research for conversion prediction (alphabetically ordered)

As shown in Table, nearly all prior studies consider features belonging to the page category, whereas most of them additionally include time-related information. Monetary-related features, features capturing direct interactions with the website and demographics are still used fairly often. However, user agent is a category which has been used only to be able to combine aggregated demographic data with available clickstream data (Chan et al. 2014) but not as specific feature category. However, we will consider this feature category for our prediction task.

From a feature horizon perspective, several studies also consider the use of a broader time horizon of information by collecting data over a longer period to add historical session and purchase information (e.g., Buckinx and Van den Poel 2005; Sismeiro and Bucklin 2004). In general, the literature is almost equally divided into studies that use only information with respect to a current website visit and studies that, in addition, use historical data related to earlier website visits and purchases. The broad use of feature categories and varying time horizons supports the relevance to consider privacy-related aspects since the broader the more information used and the longer the time horizon considered, the more privacy severe an approach might be.

3.2. Conversion Prediction and Privacy

We will next take a detailed look at all papers considered and examine those which are closely related to our work in that they raise privacy issues in combination with assessing the linked business value.

The second column *Privacy Focus* in Table 1 depicts whether prior work on clickstream-based conversion modeling makes reference to user privacy. Of all papers considered, only two papers raise privacy issues at all. Moreover, while some studies examine the link between the accuracy of a behavior prediction model and its business performance (see column *Profit / Business Value*), the potential trade-off between performance and privacy has eluded research. In appraising this result, it is important to note that some studies assess predictive accuracy using the lift measure. Although they do not investigate the business performance of their models (e.g., in terms of costs and revenues), it is possible to relate lift, under certain assumptions, to profitability (Masand and Piatetsky-Shapiro 1996). To acknowledge that at least an implicit link to business performance exists in corresponding work, we highlight usage of the lift measure in Table 1.

This section will present the two studies which are closest related to this work in that they raise the issue of privacy, which are Padmanabhan et al. (2006) and Stange and Funk (2015). The former consider the privacy aspect of user data in terms of the trade-off of using a single data source compared to using data collected across several websites. Cross-site data provides a more comprehensive picture of user behavior. However, such data is normally only available via acquisition from third-party vendors. To that end, the authors define features that are either user- or site-centered. While site-centered data only uses information from a single data source, i.e. one website at a time, and is therefore more privacy preserving, the user-centered approach captures the behavior of website visitor across all websites in their dataset. The authors investigate the prediction accuracy of their tree-based model with regard to three different dependent variables: conversion during the session, conversion during any consecutive session, and return website visit. Using the lift measure and predictive models based on all available data, the authors show that the user-centric approach always outperforms the privacy-friendly site-centric approach. However, since third-party data is expensive to obtain, it is often not a sensible option for e-commerce websites to have complete information for all visitors across websites. A fraction of 45 percent of user-centric data is necessary to build a model which is able to outperform a site-centric model trained on all available data. Therefore, including only a small extent of privacy adverse information on browsing behavior across several websites might reduce prediction accuracy compared to using comparatively privacy friendly single site data only.

Stange and Funk (2015) examine how sample size affects the predictive accuracy of a clickstream prediction model. This relates to privacy in that gathering larger samples requires companies to collect data over longer horizons, and thus act in a relatively more privacy adverse manner. Using one-month data of two online retailers, they find that including only one percent of all available clickstream data is already sufficient to predict the likelihood of conversion with satisfying accuracy. Despite looking at privacy from the perspective of the amount of data needed, their dataset still contains several privacy-harming features such as the link between advertisement and website interaction of a single user.

While existing papers with a relation to the privacy aspect of clickstream data collection focus either on the amount of data needed (Stange and Funk 2015) or whether collecting data across multiple websites exhibits benefits (Padmanabhan et al. 2006), we focus on understanding what kind of data from a single data source is sufficient to obtain accurate conversion predictions. Therefore, our research contributes to the existing literature in three ways. First, we define privacy categories for site-centric clickstream data. Second, we investigate the incremental benefits of successively including more privacy adverse data into a predictive model to understand the informational gain of the identified categories. Third, since the collection and usage of clickstream data is a business decision, we consider a specific use case to analyze the monetary value of the different privacy-relevant feature subsets.

4. Methodology

This chapter discusses the construction of the different clickstream feature sets based on the features' risk to data privacy and clarifies our predictive modeling methodology.

4.1. Definition of Privacy Settings and Feature Extraction

To grasp the connection between clickstream data, its utility for website owners, and threat for user privacy, we categorize clickstream features into groups according to the severity with which they might invade privacy. The categorization is based on the time horizon and user-centricity of the data during site visits. In general, privacy risk increases with the amount and dimensions of data, which, in turn, increase with the time horizon over which a visitor is monitored. For example, gathering clickstream data for a specific visitor over one session is less severe than monitoring this visitor's behavior for multiple sessions. Surveys show that around 69 percent of Internet users do not delete their cookies at least on a monthly basis, making it easy to re-identify the majority of revisiting website users (Statista 2016b; comScore 2007). Hence, the horizon of clickstream data gathering is one determinant of the severity of its privacy impact.

In addition to the time dimension, privacy implications vary with the type of data being collected. This is especially relevant since we will focus on website visitors who visit a shopping website anonymously (i.e., without registration). In general, the more data is available, the more holistic the picture of a visitor and the more conclusions about future behavior can be derived from the data (Bennett et al. 2012). Therefore, the richness of data collection and information extraction is a second factor that we consider in our clickstream feature categorization.

In particular, we consider website-centric data as less privacy intrusive than user-centric data. The former is related to information such as the webpages a user has visited, whereas we define user-centric data as information related to user agent and page engagement in the form of basket actions, click- and scroll events. Drawing upon the two determinants of potential privacy issues, monitoring time horizon and data richness, we propose four categories of clickstream features, which we label *SessionContent*, *SessionBehavior*, *CrossSession*, and *Identifiable*. Table 2 summarizes those feature sets where we provide an indicator of privacy relevance on a high-level basis and a description of each feature set. Furthermore, we adopted the classification approach of chapter 3 and summarize the kind of information contained at the specific privacy level, i.e. features of the current and all less critical levels of privacy.

We argue that privacy benefits stem from smaller observation periods and more site-centric features. In the *SessionContent* setting, we only include page, path, category and basic time (i.e., time on a webpage and session duration) and monetary (i.e. monetary amount in basket) related information of a session. These features are based on information directly available through the browser requests and consequent page views. In other words, features belonging solely to the current session are our information baseline, which contains the least privacy invasive information.

Privacy Relevance	Setting	Description	Clickstream Feature Category					Feature Horizon	
			Page	Time	Monetary	Page Interaction	User Agent	Current Session	Across Session
Lower Site-centered	Session Content	…uses only information of the current session of a user related to page visited and time spent on page.	x	x	x			x	
	Session Behavior	…considers interactions with the website with respect to basket, click, scroll and tab switch events.	x	x	x	x		x	
	Cross Session	…contains information spanning a longer time horizon over all current sessions of the observation period.	x	x	x	x			x
Higher User-centered	Identifiable	…contains user agent related information such as IP-resolved location, screen resolution, access device and software.	x	x	x	x	x		x

Table 2. Description of our defined settings with varying privacy horizons[13] on the dimensions of clickstream feature category and time horizon.

The *SessionBehavior* setting is related to click and scroll events, basket actions and the time spent in total on different page categories (e.g. on product and shopping basket pages). The setting is defined as more privacy intrusive since the interaction of the client on the page is observed in addition to the page visit itself. Interaction with an e-commerce webpage hints at an interest for a specific product, for example via basket actions or if several click/scroll events on a specific page signal a strong interest in the information displayed on that page. Research has shown that mouse movements map to a certain extent the gaze movement of a website visitor therefore hinting at the relevance of a specific page (Guo and Agichtein 2010a; Rodden et al. 2008). Furthermore, website interactions in form of click and scroll events have been shown to provide a stronger signal with respect to a purchase intention compared to only using content-based information (Guo and Agichtein 2010b). In addition, website interaction information can be used for re-identification purposes (O'Connell and Walker 2014). Nevertheless, privacy risks are reduced by the time frame, which is still restricted to one session.

The *CrossSession* setting contains features related to all preceding site visits within a two-month period, implying that the monitoring horizon is larger compared to previous settings. Tracking a user over multiple sessions implies that data needs to be stored over the full time period, increasing both the risk of misuse and the amount of data at risk, thus leading to a stronger privacy impact. By storing session information and connecting it via a user identifier, e.g. through cookies or browser fingerprinting, long-term user profiles can be constructed. The observation of a longer time span of user behavioral patterns of website visits and interactions give a stronger indication of the intention of the users' website visit (Bennett et al. 2012). Additionally, this enables user profiling and in the end to match the behavioral patterns over time with the identity of the website visitor (Yang 2010). Here, the sequences of and time spent on websites visited by a user is used to detect re-occurring patterns in user navigational behavior

[13] A detailed overview of the features contained in each setting can be found in the appendix.

so that yet unknown web session can be assigned with a certain probability to a specific user. This approach is even more intrusive since it can refrain from traditional tracking techniques in the form of cookies, where the website visitor can control and hinder the tracking attempts through the regular deletion of cookies from the system. Instead, via the large-scale observation of user behavior over time, sessions can be matched to particular website visitors without the definitive need of technological identifiers. Features in this category include, for example, aggregates such as the overall number of page views of a single user, her mean time spend on a page, or differences in interaction patters of the current visit compared to previous visits (e.g., time on page compared to this user's mean time on page).

Defined as our most privacy intrusive setting, the category *Identifiable* contains user agent information such as IP-resolved location and details related to browser, access device and screen resolution. Clickstream data in itself can be collected anonymously and restricted to the tracking of a user within one session, e.g. by the deletion of cookies. To facilitate behavioral pattern matching for user identification, a certain observation horizon is necessary to obtain reasonable results (Yang 2010). However, browser fingerprinting can be used to recognize and track online users by the setup of their system (e.g., Eckersley 2010; Boda et al. 2012), since it is transmitted automatically with a page request. Features of this setting include data on the customer's access infrastructure, e.g. device type, device brand, operating system, and browser, location and update recency of the user's system. They also include the IP address and the inferred location of the user. Since this data can be used to locate and identify the user in context, we classify these features as potentially personally identifiable information and most privacy concerning.

To facilitate the prediction of user behavior, raw clickstream data is converted into clickstream features. In total, we derive a set of 84 clickstream features. The full list of features used can be found in the appendix.

4.2. Predictive Modeling

This chapter describes our predictive modelling approach in terms of algorithms used and the specific set-up to clarify how our derived clickstream feature sets influence predictive accuracy. Not gathering any clickstream data might be most desirable from a privacy perspective. However, this clearly conflicts with website owners' business goals and their interest to gather data for user behavior prediction. To clarify the trade-off between respecting user privacy and collecting informative data, we approximate the value of clickstream data in a predictive modeling context.

We train prediction models to estimate the purchase probability for each visitor in the current session after each click given the features described above. This method is known as "clipping at every click" (Van der Meer et al. 2000; Senécal et al. 2014). More specifically, we aim at predicting whether an e-shop visitor will make a purchase in her current session (e.g. Padmanabhan et al. 2001), where one user session comprises multiple page views. If a purchase is made at any point within a session, we define the target variable to be positive for all page views in that session; and as negative otherwise. We then estimate a statistical model that classifies each new observation, i.e. each page view of a user on the website, into one of the two categories "user will purchase during this session" and "user will not purchase during this session". We predict purchases at the page view level since marketing stimuli like e-coupons can be offered at any point in a session and therefore require page-level granularity.

In formal terms, we face a binary classification problem with groups purchase/no purchase. Several machine learning algorithms are available to estimate classification models (e.g., Lessmann and Voß 2009) and the analysis of model performance and variable importance is dependent on the model choice. For the purpose of this study, we select the random forest algorithm due to its prevalence in practice and good track record in many applications (Kuhn and Johnson 2013) and because we observe it to perform best in terms of overall prediction error when compared to other models on our data. We determine model performance by pretests comparing the predictions of random forest, C5.0 and gradient boosting including parameter tuning on the full data.

Random forest is an ensemble algorithm. It combines hundreds of decision trees build on subsamples of the observation and feature space to ensure diversity among individual trees. Each tree is a sequence of binary splits of the data that maximize class purity in leaf nodes. For each observation, the random forest model estimates the probability of it to belong to class *purchase* by the ratio of trees that predict this class.

To build the model and test it on unseen observations, we split our data into a training and test set consisting of data from August and September 2015, respectively. We estimate the model from the training data and use the test data to assess predictive accuracy. Prior to that, we perform 5-fold cross-validation on the training data to identify suitable values for the algorithm parameters (i.e, the number of decision trees in the random forest and the number of randomly selected variables per tree split), as is standard practice in predictive modeling (e.g., Kuhn and Johnson 2013).

Using the above approach, we examine how the accuracy of purchase prediction models varies with the clickstream features they embody. Therefore, model selection, training, and testing are conducted separately for each of the clickstream feature subsets. More specifically, we consider two experimental settings. First, the models are trained on each feature subset in isolation. Results of this setting provide an estimate of the predictive value contained in the underlying features. Second, we train models on an incrementally increasing set of features, where more privacy adverse features are added in each step. For example, we start with training a model using only the least sensitive features of the *SessionContent* setting. Next, we add the features of the *SessionBehavior* setting and train a second model. We continue the incremental addition of feature subsets in the order of privacy adverseness until all feature subsets are considered. Taking both the incremental expansion of feature sets and the test of feature sets in isolation into account, we assess random forest models based on a total of seven distinct feature sets (i.e. four times each individual subset plus three times incrementally developed subsets), each with a different number of features and degree of sensitivity with regard to customer privacy.

This modeling approach allows us to quantify the marginal value of adding features of a presumably more comprehensive, but also more privacy adverse category. However, we also acknowledge a limitation of our approach, namely that it disregards any additional value that clickstream features and gathering the corresponding raw data, respectively, provide to the website (i.e., shop) owner beyond facilitating predictive modeling. Our justification for concentrating on predictive modeling is twofold. First, analysis of the user journey and product-centered browsing behavior can largely be performed on aggregated clickstream data. Thus, gathering clickstream data at the individual user level is likely dispensable for strategic site management tasks. Second, the individual purchase and personal information used in most business intelligence applications is provided willingly by users upon registration or purchase. Unlike the anonymous visitors who we focus on in this paper, registered customers have explicitly agreed to further data collection.

5. Empirical Results

Adopting the methodology discussed above, we analyze the performance of the classifier for the defined subsets of features in two ways. First, we employ statistical performance measures to create a comparable benchmark of the general predictive power of each feature set. To complement the accuracy assessment, we approximate the economic value of different models. While being specific to one application context (e.g., specific cost and revenue consideration) and thus less general, we consider the economic analysis to add useful insight from a managerial perspective.

5.1. Data and Data Preparation

As a basis of our study, we obtain clickstream data from two large European online retailers of two e-shops selling apparel and shoes, respectively. Both shops are comparable in size as measured by the number of users, sessions and views within the two-month observation period as described in Table 3. The clickstreams include desktop and mobile users who accessed the shop websites in a two-month period from August to September 2015. As part of data preparation, we exclude the first three page views within each session, since we assume constant coupon success probability throughout the profit analysis (see Section 5.4.). An empirical analysis of the data shows very high exit rates for these views suggesting that a large number of visitors does not enter the page with a strong intention to interact. In case of both shops, around 55 percent of website visitors leave the website after only three webpage views. High exit rates further suggest a strong dependency between the current view and coupon success probability, which would require explicit modelling of redemption rates for the profit simulation beyond the scope of this paper. From business perspective, coupon marketing on the first three page views amplifies the number of played coupons at a very low redemption rate. This is generally not in line with company expectations due to concerns about customer price expectations regarding the availability of coupons and the brand image.

Based upon the empirical analysis of the distribution of the length of sessions, we further deleted sessions with more than 500 page views each under the assumption that such sessions come from bots (Banerjee and Ghosh 2001). However, this affected only one session in both datasets. Furthermore, since we include clickstream features that span several user sessions, we select customers with at least four visits during the two-month period. This is to ensure that all experimental settings have access to the same observations. Table 3 summarizes the resulting data, which includes 120,554 unique user sessions from 18,852 customers with 1,520,418 page views in total. The overall conversion rate is about 6.11 % (7,362 sessions).

Summary	Shop 1		Shop 2	
	Tablet	Computer	Tablet	Computer
Users	2,055	9,247	1,463	6,087
Sessions	10,947	51,349	11,171	47,087
Views	120,845	585,570	182,726	631,277
Purchases	744 (6.80%)	4,112 (8.01%)	538 (4.82%)	1,968 (4.18%)

Table 3. Summary of the datasets of both shops used in the empirical study

5.2. Analysis of Predictive Performance

In this section, we analyze the predictive value of each feature set. On the basis of the clickstream features corresponding to an observation (i.e., a page view), the random forest model estimates the probability that a purchase will be made in the corresponding session. Comparing this probability estimate to a cutoff, one obtains a discrete classification of observation into the two groups purchase/no purchase. It is common practice to assess the accuracy of classification models using receiver operating characteristic (ROC) analysis.

Without any further assumptions about the application setting, we report the value of the feature sets measured by the area-under-the-ROC-curves (AUC) of each model build on the respective features, where an AUC of 0.5 and 1 indicate the performance of random assignment and perfect separation, respectively (Kuhn and Johnson 2013). Comparing the feature sets based on model performance has the advantage to capture any redundancy or interaction effects between features of different subsets. Tbale 4 (right) shows the marginal improvement of clickstream features via a stepwise extension adding a set of more privacy adverse features at each step. For both shops, we observe that extending the *SessionContent* feature set by including page interaction features does not substantially improve the predictive model. Starting with an AUC of 0.797 for Shop 1 and 0.759 for Shop 2 in the *SessionContent* setting, the inclusion of behavioral information only accounts for an improvement of 0.004 (Shop 1) and 0.006 (Shop 2), respectively. We tentatively conclude that there seems to be little if any (predictive) value in combining the two sets of clickstream features in this behavior prediction model. In contrast, extending the feature set by *CrossSession* features increases performance by about 0.03 AUC points. Given the baseline level of AUC equal to 0.80 and 0.76 for shop 1 and shop 2 respectively, an increase of 0.03 may signal a sizeable improvement of model performance in economic terms. Finally, the features concerning user and system information appear to not add any predictive information beyond that already embodied in the random forest model of the previous step. This follows from the, once again, very small performance increase of about 0.002 AUC points (both shops). At the same time, this data is most likely to reveal a user's demographic information or identity, thus bearing the highest potential risk to user privacy.

The AUC performance for the features sets separately presented in Table 4 (left), indicates to what extend more privacy adverse features contain information already captured by less invasive variables. In line with marketing intuition, characteristics representing information originating from a single session and on-page behavior within a session are the strong predictors of customer conversion (Chaffey 2015). It is worth noting that *SessionContent* information, identified as the least privacy adverse type of customer information, matches and outperforms both more privacy intrusive feature sets and long-term profile data on this metric. Indeed, the aggregation of session and behavior data in the form of cross-session features, which can be used to create a customer profile, is substantially less informative on its own for shop 1, while being slightly higher for shop 2. The weak predictive power of the *Identifiable* set in the incremental setting is mirrored when considering the set on its own. Apparently, the data on the user agent header with regard to information on a user's location and system does not show substantial predictive power as indicated by an AUC value of just above 0.5. We will come back to these findings in Section 5.3 when discussing the importance of individual variables.

	AUC			
Feature set	**Sets separately**		**Incremental extension**	
	Shop 1	Shop 2	Shop 1	Shop 2
SessionContent	0.797	0.759	0.797	0.759
SessionInteraction	0.781	0.758	0.801	0.765
CrossSession	0.760	0.763	0.832	0.801
Identifiable	0.535	0.528	0.834	0.803

Table 4. AUC values for predictive models build on the feature sets separately (left) and the incrementally increasing feature set (right)

In summary, the analysis of predictive accuracy using the AUC provides evidence for our datasets that simple features based on page information within the current session are sufficient to achieve a performance level close to using the full set of clickstream features considered in the study. On the other hand, a notable performance increase over the baseline setting using only the least privacy adverse features has been observed for our data when adding *CrossSession* features.

5.3. Importance of Individual Variables

Before we go on to analyze the effect of the observed performance gains in terms of monetary profit, we analyze the importance of the variables to the model individually in order to identify the main predictors in each of the feature subsets. This allows us to further differentiate the marginal benefit of collecting very specific data with the potential to 1) reduce the number of sensitive features to be collected by focusing on a (small) subset of important predictors and 2) develop less sensitive proxy data for important predictors. At the same time, this section extends research to identify important predictors of customer purchase behavior. This analysis is limited by the dependence of the importance score on the random forest algorithm used to build the model and evaluate the feature importance.

The random forest algorithm provides a measure of relative variable importance, which captures the degree to which corrupting a variable decreases the predictive performance of the classification model (Breiman 2001). To determine the importance rank of a variable, random forest calculates the classification accuracy of each individual decision tree using the observations not employed for growing this tree, adulterates the variable by adding random noise, re-assesses the component trees' accuracy, and averages the difference in accuracy before and after variable corruption across all trees in the forest. The larger the decrease in accuracy, the more important the variable.

Figure 2 shows the variable importance ranks for the 25 most predictive variables in the best-performing random forest, which is based on the full feature set. The variables are ordered according to their average relevance for both shops, i.e. the averaged importance values for each variable of both shops. Importance estimates are normalized in the range of 100 to 0 indicating maximal and minimal importance, respectively. We use stars to identify the membership of a variable to one of our four feature sets.

Overall, variable importance develops consistently across shops. Notable differences can be observed for features that capture information concerning basket or checkout interaction. The corresponding clickstream features belong to the *SessionContent* and *SessionBehavior* set, which the previous analysis has shown to encompass similar information. In other words, features can substitute each other, so that variation in the importance ranking across shops is plausible. Pearson correlation scores for the discussed features support this interpretation and are included in the appendix. Correlation between variables may also impact the importance scores of the correlated variables by mitigating (acerbating) the accuracy decrease resulting from permutation in case of positively (negatively) correlated variables (Gregorutti et al. 2017). For the results of Figure 2, the correlation patterns (see Figure A in the Appendix) suggest that random forest importance scores might underestimate the actual relevance of the top three features due to positive correlation, whereas importance scores of features describing the operating system, the purchase recency and the time since adding a product to the basket, which are negatively correlated with some other features, might have been overestimated. However, in view of a relatively large forest size of 700 trees, such effects, it they exists, are likely to be small and should not distort overall tendency in feature importance.

Figure 2 provides three main insights. First, the *SessionContent* features account for six of the ten most important features. Within these, features describing the time of visit and an interaction with the shopping cart are most predictive. This result comes with the caveat that basket-related features are informative only after interaction with the basket has taken place, although it is important to note that this interaction includes viewing or removing products from the basket during search phase. The fact that the most important features of the *SessionContent* setting and the *SessionBehavior* setting both convey information related to shopping cart interactions also explains why the inclusion of *SessionBehavior* features does not significantly improve prediction performance (see Section 4.2).

Second, variable importance is highly skewed overall and within each subset. This suggests that it might be possible to reduce the number of features and thus the amount of data being collected about shop visitors without sacrificing predictive accuracy. The *CrossSession* features are an exception, which make up the body of important features at a rather low relative importance, but have been found to significantly increase predictive performance when they are included as a set. This suggest that there is no small subset of *CrossSession* features that could be singled out to provide the performance gain while less privacy-sensitive features are collected.

Third, we find information on user location and device size to be important, although they have not increased predictive performance in Section 5.2. This suggests that the predictive information contained in these features is also embodied in clickstream features from other sets.

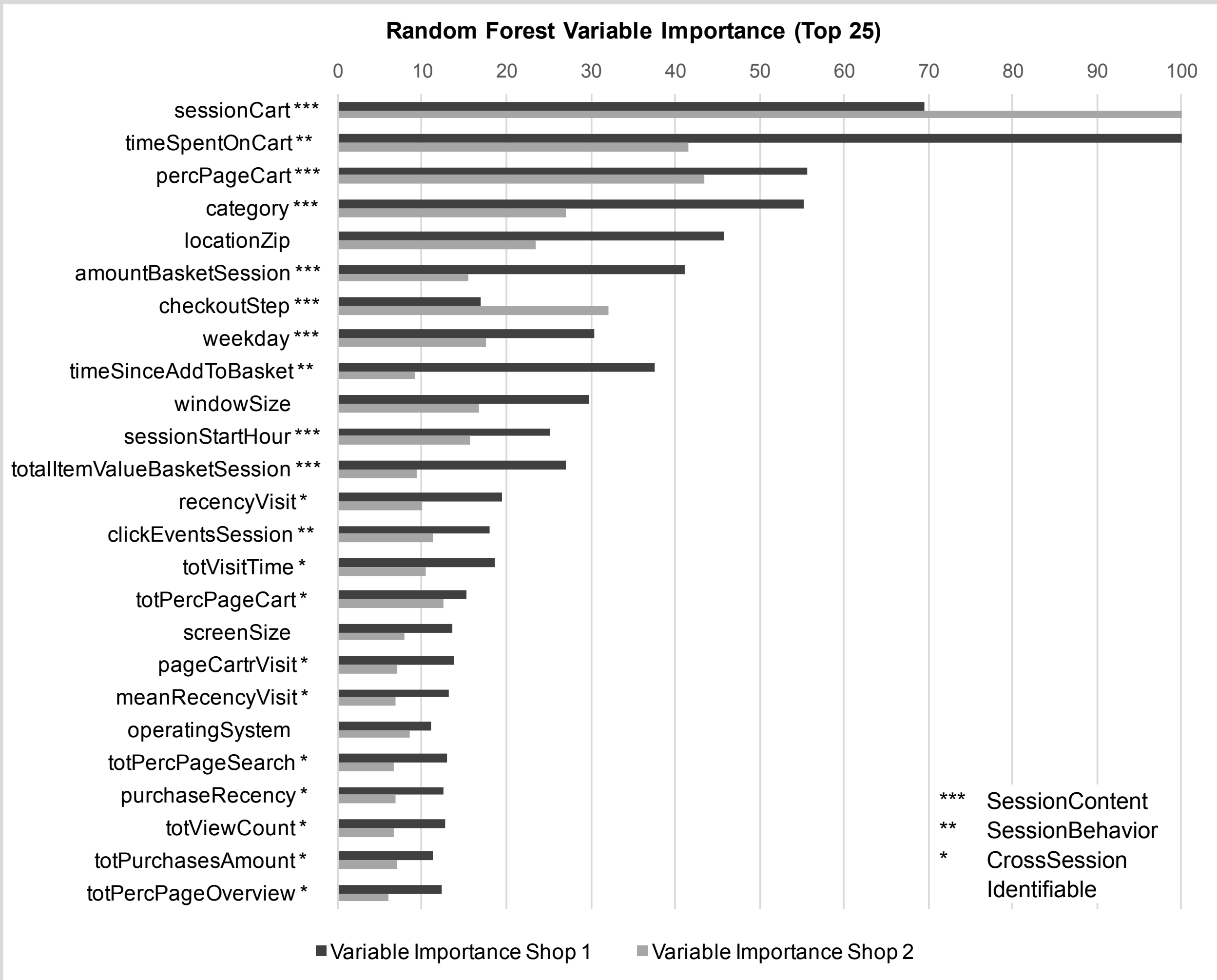

Figure 2. *Random forest variable importance for the 25 most important variables ordered according to their average relevance for both shops.*

An important qualification to these results is that the information captured in *SessionContent* and *SessionBehavior* is created over the course of the session and is therefore only available at later stages. This is different for features of *Identifiable* which are readily available at the very beginning of a session and also for features of *CrossSession* in case of returning visitors. This restriction is particularly relevant for applications where marketing contact is fixed to a specific page view before or at which the prediction must take place. From a data perspective, providing an incentive to reduce basket abandonment is potentially profitable even at late stages of the purchase process. Table 5 shows the ratio of customers in our dataset that do not complete their purchase after each page in the purchase process.

Here, in case of shop 1 four steps are necessary for purchase completion whereas in case of the shop 2 only three steps are required. When finalizing the purchase process, customers undergo steps such as reviewing basket contents, entering their shipping information and the final confirmation of their purchase. Even at this point and at the last purchase step, abandonment rates are still as high as 13% and 25% for shop 1 and 2, respectively. High basket abandonment rates can be caused by a sudden change in customer intention potentially acerbated by unintuitive website design or (lack of) shipping and payment options.

	Purchase step				
	0	1	2	3	4
Shop 1	89%	61%	38%	24%	13%
Shop 2	86%	58%	39%	25%	-

Table 5. *Basket abandonment rates (in %) for each step in the purchase process (four steps in case of shop 1, three steps in case of shop 2).*

5.4. Economic Value of Customer Data

Statistical measures of predictive performance and variable importance avoid assumptions on the application setting and thus represent a universal indicator of predictive power. This advantage is also a downside. In particular, an interpretation of the AUC or a variable importance score might not capture the characteristics of a specific application context. Moreover, the performance indicators used in management practice typically comprise measures of economic values. In this sense, managers might find it difficult to appreciate a difference in terms of the AUC and make decisions on the ground of such information. More specifically, the results of Section 4.2. indicate that features belonging to *SessionInteraction* and *Identifiable* setting are irrelevant for prediction. This suggest that there is no need to gather corresponding data. Likewise, including *CrossSession* features has been found to improve accuracy, which implies that the e-shop should continue to collect user information across multiple sessions. However, the consequences of these decisions remain abstract when examined in the dimension of AUC differences. A cost-benefit-analysis, although being less general, provides useful additional information for managerial decision making. We therefore simulate a specific business scenario, namely coupon targeting, in order to analyze the monetary value associated with the use of different clickstream feature subsets. This achieves two goals. First, it provides a realistic reference value regarding the business value of sensitive customer information, and second, it outlines the process that is required to express the question of data collection in monetary terms and to make informed business decisions.

To pursue these goals, we consider the marketing context associated with the data and assume that the e-shop strives to increase sales by means of couponing. E-coupons are dynamically incorporated into a webpage and thus each user's session and have gained substantial popularity to stimulate purchases in e-commerce (e.g., Khajehzadeh et al. 2014). When a coupon is offered to a visitor who is not inclined to buy, there is a probability p that she will purchase, which we assume to be constant over users. If a purchase takes place the e-shop receives expected revenue of r reduced by the cost of the marketing incentive c, where generally $c < r$ by design. However, when a coupon is offered to a customer who would buy naturally, the company faces an opportunity cost equal to the coupon value c. Assuming no other strategic restrictions on coupon offerings apply, it is optimal to offer a coupon to all those and only those customers, who do not plan to purchase naturally, as identified by the classification model. The cost-benefit matrix for the setting considered here (Table 6) has, to the best of our knowledge, not been described in previous literature, but differs from the standard coupon targeting setting only in so far as the cost associated with the coupon is realized only when a purchase takes place.

<table>
<tr><td colspan="2"></td><td colspan="2">Actual Decision</td></tr>
<tr><td colspan="2"></td><td>Purchase planned</td><td>No purchase planned</td></tr>
<tr><td rowspan="2">Predicti</td><td>Purchase / No coupon</td><td>r</td><td>0</td></tr>
<tr><td>No purchase / Issue coupon</td><td>r - c</td><td>p·(r - c)</td></tr>
</table>

Table 6. Campaign revenue matrix for the coupon campaign setting.

We can express the net revenue matrix (Table 6) in the form of a decision-equivalent cost matrix (Table 7), where the costs on the diagonal are normalized to be equal to zero without an effect on the optimal probability threshold (Margineantu 2001; Elkan 2001). This cost matrix better expresses the optimization problem faced by the decision model. The model aims at distinguishing purchasers from non-purchasers under the constraints that 1) issuing a coupon to a purchaser unnecessarily reduces sales profit by the coupon value c and 2) not targeting a non-purchaser foregoes a chance to convince the customer, which is associated with an opportunity cost of the expected sales value.

<table>
<tr><td colspan="2"></td><td colspan="2">Actual Decision</td></tr>
<tr><td colspan="2"></td><td>Purchase</td><td>No purchase</td></tr>
<tr><td rowspan="2">Predicti</td><td>Purchase / No coupon</td><td>0</td><td>-p·(r - c)</td></tr>
<tr><td>No purchase / Issue coupon</td><td>-c</td><td>0</td></tr>
</table>

Table 7. Derived cost matrix for the coupon campaign setting.

The performance of a classifier in monetary terms then depends on its ability to distinguish accurately between actual buyers and non-buyers (i.e., classification accuracy), the ratio between r and c, and the success probability of the coupon, p (i.e., conversion rate). The dependence on parameters such as c, r, and p explains why an economic evaluation of a predictive model is less general than an evaluation based on the AUC. We compute total sales revenue by multiplying the number of customers in each class with the respective revenue for the class.

In the following, we set basket revenue r to the average basket value observed in our data, which is 54.37€ and 49.45€ for shop 1 and shop 2, respectively. For c, we select a 10% reduction on the basket value approximately matching the 5€ coupon value employed by the respective shops in their campaigns, which we consider consistent with general marketing practice. The average face value of online coupons in the non-food area has been shown to be around 2€ in 2016 (KantarMedia 2016) indicating that our approach is more pessimistic in terms that the wrong classification of a non-buyer as buyer yields to a more severe punishment, i.e. a higher financial loss.

In addition to discount values, coupon conversion rates (i.e. how many customers accept a coupon and complete a purchase) are likely to depend on industry, online shop and product category characteristics as well as other criteria. To the best of our knowledge, prior literature does not offer insights concerning average coupon conversion probabilities across these categories. Likewise, publicly available information on this matter is limited, which is intuitive considering that corresponding information is sensitive. Some evidence is available for China where most successful websites achieve conversion rates up to 6% (Statista 2017). However, this data comes from 2011 and does not distinguish between coupon types, values, industries, etc.

In the interest of generality and to ensure robustness of results, we therefore consider several coupon conversion rates p between 1% and 5% and simulate the business value of a coupon targeting model for these settings. The choice of the conversion rate interval centers around the global conversion average for online shoppers of 2.5% (Statista 2016c). The interval is also consistent with (Statista 2017). Note that savings associated with the correct identification of a buyer stay constant over coupon conversion rates, while the cost of misclassification increases with coupon success probability.

From the conversion rate shown in the dataset description (Table 3) and the cost ratio given in Table 7, it is clear that the prediction problem is imbalanced, i.e. that the non-purchase class is more common than the purchase class, and cost-sensitive, i.e. that the misclassification of purchasers as non-purchasers is more costly than vice versa. To account for both issues, we apply a post-processing method for each feature set and choose the revenue-optimal probability threshold empirically on the training data (Sheng and Ling 2006). To obtain discrete class assignments from the random forest classifier, which produces purchase probabilities, we compare probabilistic predictions to a threshold and classify users as buyers if the random forest predicts a purchase probability above the threshold; and non-buyers otherwise. By setting a higher (lower) threshold, less (more) users are classified as purchasers and receive a coupon, thus adjusting for the class distribution and cost setting. We select the revenue-maximal threshold for each feature set and coupon effectiveness by calculating the revenue on the training data for a range of thresholds in [0;1]. This way, we identify the threshold that leads to the highest revenue for each model and use this threshold when applying the model to classify the users in the test set.

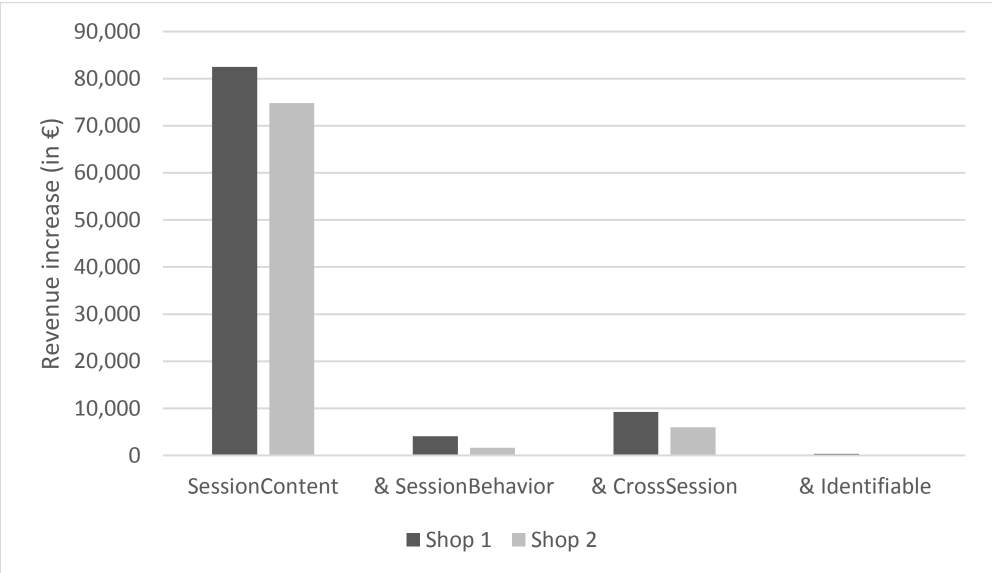

Figure 3. *Revenue increase as compared to campaign without customer targeting (all/no coupons), averaged over coupon success rates from 1%-5%.*

Given these assumptions, Figure 3 shows the net revenue generated over 247,325 and 251,786 customers in the test set for shop 1 and 2, respectively, by employing a customer targeting model based on each of the feature sets averaged over the range of coupon success rates. We consider as benchmark the revenue of a no-model solution, i.e. a hypothetical campaign where either no or all customers receive a coupon, whichever is more profitable given the respective coupon success rate. We calculate the revenue gain of the decision model by subtracting the revenue of the benchmark from the total model revenue. A substantial average increase in revenue of 82,482€ and 74,792€ for shop 1 and shop2, respectively, is generated by the predictive model employing *SessionContent* features. Additional gains achieved by the inclusion of *SessionBehavior* and *CrossSession* features are comparatively smaller at below 5,000€ and 10,000€, respectively. The overall revenue of the campaigns and the net revenue gain of each feature set compared to the next less sensitive set for each coupon success rate, which is the basis for Figure 3, are reported in Table 8.[14]

		Total campaign revenue (in 1000€)					**Net gain to less sensitive set (in 1000€)**				
Conversion rate		**1%**	**2%**	**3%**	**4%**	**5%**	**1%**	**2%**	**3%**	**4%**	**5%**
Shop 1	***Naïve: All/No Coupons***	2612	2612	2612	2694	2780	**-**	**-**	**-**	**-**	**-**
	SessionContent	2629	2679	2736	2804	2874	17.2	66.6	124.5	109.8	94.3
	& SessionBehavior	2629	2683	2743	2808	2879	0.2	4.5	6.8	4.3	4.8
	& CrossSession	2633	2690	2753	2821	2892	3.6	7.1	9.6	12.9	13.1
	& Identifiable	2633	2690	2753	2821	2893	0.1	-0.3	0.3	0.7	0.8
Shop 2	***Naïve: All/No Coupons***	2665	2665	2720	2820	2921	-	-	-	-	-
	SessionContent	2668	2744	2829	2917	3007	2.9	78.8	108.9	96.8	86.6
	& SessionBehavior	2667	2746	2831	2920	3010	-0.8	1.9	1.7	2.4	3.1
	& CrossSession	2680	2752	2835	2923	3014	13.2	6.3	3.8	3.0	3.6
	& Identifiable	2679	2751	2835	2924	3015	-1.0	-1.2	0.4	1.3	1.4

Table 8. Simulation results for asymmetric cost in terms of total campaign revenue (left) and relative gain compared to the next less sensitive data subset (right)

The results provide two main insights. First, substantial cost savings can be achieved by better coupon targeting using the least privacy invasive feature set. Compared to a hypothetical benchmark campaign, where either no coupons are handed out or all customers receive a coupon, the savings amount to between 65,000€ and 125,000€ per month for all but the 1% coupon success rate scenario, which are realized by targeting only customers with a low conversion probability or excluding expected buyers from the coupon campaign, respectively. Even when coupons are assumed to be least effective at p = 1%, the most basic *SessionContent* model creates savings of 24,000€ and 3,000€ for shop 1 and 2, respectively. The high gains indicate that the collection of session data and the development of a predictive model are highly profitable in this example.

Second, making use of more sensitive customer data does not lead to a linear increase in campaign results. The marginal gain from *SessionBehavior* features lies between 1,000€ and 7,000€ with the average at approximately 4,500€ and 2,000€ for shop 1 and 2, respectively. The addition of *CrossSession* features entails an observed revenue gain of between 3,000€ and 13,000€ for shop 1 and 2, respectively. In four cases, the addition of features results in a small observed revenue loss, particularly for the *Identifiable* features. As apparent from the AUC (see Table 4), the extended model does not in fact provide worse predictions. However, the model and its predictions are so similar to the less invasive feature set that slight variation in the empirically tuned probability threshold cause slightly better or worse performance on the test set. The generally insubstantial difference in performance of the *Identifiable* set to the *CrossSession* set undermines the naïve credo that more data is always better.

In summary, while there are monetary gains achievable through the employment of more sensible features for coupon targeting for this data and application, the largest part of the realized gains is achieved with comparably privacy-friendly data. In all but the 1% redemption rate scenarios, the most basic *SessionContent* data allows the realization of at least 90% of the highest feasible savings, not adjusted for the costs of data collection, storage and protection, and additional risks that come with the handling of more sensitive data. The simulation in Table 8 thus facilitates the conclusion that the

[14] The calculations are based on the actual number of correctly and incorrectly classified customers across the 50 (2 shops x 5 feature sets x 5 conversion rate) settings. Interested readers find results at this level of detail in the Appendix.

collection and application of a sensitive set of customer features beyond non-behavioral session data is unnecessary to achieve large returns on investment.

At the same time, the absolute gain provided by the collection of *SessionBehavior* and *CrossSession* information may be judged to be substantial from a business perspective. Expressing the net revenue difference in terms of the maximum gains obtainable by making use of the full set of features, the simulation suggests that a company in the assumed setting foregoes an average of 15% (shop 1) and 7% (shop 2) of potential revenue by refraining from collecting information more sensitive than *SessionContent* features. These numbers exclude the special case of a conversion rate of 1% for shop 2, where 80% of revenue is associated with *CrossSession* features. Especially for a repeated campaign setting, the expected gain from *CrossSession* features would have to be weighed against privacy considerations. More clearly, the set of *Identifiable* features, which we classify as most sensitive with regard to customer privacy, show no substantial advantage in predictive power and revenue gain in this simulation.

6. Conclusion

We investigate the marginal gain of employing clickstream data for purchase prediction of website visitors in relation to the risks to data privacy associated with data collection. The goals of our study are three-fold. First, we define four categories of clickstream information based on the threat to data privacy, namely *SessionContent*, *SessionBehavior*, *CrossSession* and *Identifiable* information in order of increasing risk. We use this framework to classify the features extracted from large clickstream datasets from two online retailers. Second, based upon this data we empirically analyze the marginal gain in predictive accuracy for the prediction of purchase behavior associated with using more sensitive customer information. This encompasses an evaluation of the importance of each feature and the performance of privacy-based feature sets both individually and aggregated. Third, we simulate a specific marketing application as undertaken by these retailers to estimate the monetary value of targeted marketing actions associated with refraining from using privacy adverse types of clickstream data.

Using a random forest model, we show that for the considered datasets the most privacy preserving *SessionContent* setting delivers competitive results in terms of customer behavior prediction. These results are improved by combining the data with *CrossSession* information about past site visits, whereas the collection of on-page behavior during the session represented by *SessionBehavior* and *Identifiable* user information do not significantly improve prediction performance.

In order to estimate the business value of extending the collected data, we simulate a coupon marketing campaign through which the e-commerce shops increase conversion rates by offering coupons to website visitors. The random forest model is used to optimize campaign targeting by identifying users that will purchase without the marketing incentive. The simulation confirms that *SessionContent* or *CrossSession* information provide a sizeable economic benefit for the considered e-commerce shops. In this setting, we estimate the opportunity costs of not collecting behavioral data and aggregating clickstream data over time at about 15% (shop 1) and 7% (shop 2) in terms of the maximum revenue obtainable by making use of the full set of features. These results imply some variation between shops and some space for e-commerce businesses to decide whether the costs and risks associated with data collection and storage are worth the marginal gain.

With respect to individual variables, we attribute the good performance of privacy-preserving *SessionContent* features to information about the page category, value of the current basket, and the time of the page view. Overall, more than half of the 15 most important variables are classified in the

SessionContent setting, while the second most important setting is *SessionBehavior* ranking as the second-best privacy preserving setting.

Our study also exhibits limitations that could be addressed in future work. First, there is some potential to extend the information sources considered in the feature set. We focus on site-centric data and disregard user-centered data collected over a range of websites by third-party entities, since this kind of data is costly to acquire for e-commerce shops. Extending the feature set by cross-site information would further increase the potential for behavior prediction, while aggravating the potential for personal identification of users and the misuse of their data. Future research could also extend the *Identifiable* setting by more involved data collection methods to extract information by cross-referencing IP addresses or retrieving installed plug-ins, language settings supplied and similar information provided by the browser. Likewise, focusing on the trade-off between privacy and profitability, we analyze empirical results across groups of variables with different privacy implications. Given the large number of variables, an analysis of privacy implications at the level of an individual variable seems impractical. However, such analysis would be useful from a business perspective to provide insights concerning the predictive and economic value of individual variables and inform shop owners which data to gather. For example, a comprehensive analysis of the partial dependence plots for the random forest model could provide further insights into the specific non-linear effects of each variable on the model prediction.

Second, we report model performance and variable importance at any view during the session. While our analysis of basket abandonment rates shows potential for marketing activities even at late stages of the purchase process, applications that are restricted to data collected until an early point during the session will likely observe a higher relevance of information that is unrelated to the current session. The optimal point in time to play a coupon and, somewhat related, the most effective type of coupon to be used, e.g. percentage-deduction or free-shipping, are interesting in themselves, but must be left for future analysis.

Third, we look at the monetary value of privacy preserving clickstream prediction in isolation and disregard any additional value of the collected data. While sales data and aggregated clickstream data are expected to be sufficient for standard marketing analyses, there clearly is potential for a more comprehensive value analysis. In particular, live testing in a real-world setting would be a promising approach to validate the monetary costs of restricting data usage determined in the simulation.

Appendix

Setting	#	Feature Label	Feature Description	Based on Buckinx and Van den Poel (2005):
Session Content	21	AvgAmountBasket	Average item value in basket	
		Category	Category type of current page.	X**
		CheckoutStep	Which step in the checkout step are we currently?	
		CountPagesRevisited	How many pages were visited at least twice during the session?	
		DayOfMonth	Day of the month (1 – 31).	
		PageVisitedBefore	Current page visited before in session.	
		PercPageCart	Relative amount of visited pages of type 'cart' / 'overview' / 'product' / 'sale' / 'search' in session.	X*
		PercPageOverview		X*
		PercPageProduct		X*
		PercPageSale		X*
		PercPageSearch		X*
		SessionCart	Number of visited pages of type 'cart' / 'overview' / 'product' / 'sale' / 'search' in the session.	X*/**
		SessionOverview		X*/**
		SessionProduct		X*/**
		SessionSale		X*/**
		SessionSearch		X*/**
		SessionStartHour	Hour of the session start (morning - midday - evening - night).	
		SessionTime	Session duration of this session.	X
		TimeOnPage	Time spent on page.	
		ViewCountSession	Number of current page view during current session.	X**
		Weekday	The weekday the session was started (1 – 7).	
Session Interaction	21	ClickEventCart	Number of click events on 'cart' / 'overview' / 'product' / 'sale' / 'search' pages relative to the total number of click events in session.	
		ClickEventOverview		
		ClickEventProduct		
		ClickEventSale		
		ClickEventSearch		
		ClickEventsSession	Number of total click events during session.	
		ScrollEventCart	Number of scroll events on 'cart' / 'overview' / 'product' / 'sale' / 'search' pages relative to the total number of scroll events in session.	
		ScrollEventOverview		
		ScrollEventProduct		
		ScrollEventSale		
		ScrollEventSearch		
		ScrollEventSession	Number of total scroll events during session.	
		TabSwitchCart	Number of tab switches on 'cart' / 'overview' / 'product' / 'sale' / 'search' pages relative to the total number of tab switches in session.	
		TabSwitchOverview		
		TabSwitchProduct		
		TabSwitchSale		
		TabSwitchSearch		
		TabSwitchSession	Number of total tab switches during session.	
		TimeSinceAddToBasket	Time passed since the most recent product was added to basket.	
		TimeSpentOnCart	Relative amount of time on pages of type 'cart'.	
		TimeSpentOnProduct	Relative amount of time on pages of type 'product'.	
CrossSession	26	ConvertedBefore	Did the client buy something in a previous session?	X
		CountPagesRevisitedLastSession	How many pages are visited again in this session compared to the former session?	
		CurrentPageVisitedLastTime	Has this specific page been viewed during the last visit?	
		CurrentViewVsPreviousAvg	Ratio of current view count relative to average view count of all previous sessions.	
		CurrentVisitLengthVsAvg	Ratio of current session length relative to average length of all previous sessions.	
		ExitEqualLanding	Is the landing page of this session equal to the exit page of last visit?	
		FrequencyVisit	Total number of recorded sessions.	X
		Hurry	Is the average time between clicks in this session shorter compared to the average time between clicks for all recorded sessions?	X
		MeanRecencyVisit	Average number of days between visits.	X
		PageCart/Visit	Number of visited pages of type 'cart' / 'overview' / 'product' / 'sale' / 'search' relative to total number of visits.	X*
		PageOverview/Visit		X*
		PageProduct/Visit		X*
		PageSale/Visit		X*

		PageSearch/Visit		X*
		Purchases/Visit	Number of purchases relative to number of visits.	X
		RecencyPurchase	Number of days since the last purchase.	X
		RecencyVisit	Number of days passed since the last visit.	X
		TotPercPageCart	Relative amount of visited pages of type 'cart' / 'overview' / 'product' / 'sale' / 'search' for all recorded sessions.	X*
		TotPercPageOverview		X*
		TotPercPageProduct		X*
		TotPercPageSale		X*
		TotPercPageSearch		X*
		TotPurchasesAmount	Total amount of purchases during all recorded sessions.	X
		TotPurchasesItems	Total amount of items purchased during all recorded sessions.	
		TotViewCount	Total number of views / clicks for all session.	X
		TotVisitTime	Total visit time for all session.	X
Identifiable	16	Browser	The type of the browser the client uses.	
		BrowserVersion	Does the client have the most recent browser version?	
		Device_Comp	Is the access device a desktop computer?	
		Device_Tab	Is the access device a tablet?	
		LocationCity	The city the client accesses the website from.	
		LocationCountry	The country the client accesses the website from.	
		LocationZip	The zip code area of the city the client accesses the website from.	
		MajorCity	Does the client access the website from a major city?	
		OperatingSystem	The operating system the client uses.	
		OperatingSystemVersion	Does the client have the most recent version of the operating system?	
		ScreenHeight	The screen height resolution of the client.	
		ScreenWidth	The screen width resolution of the client.	
		TabVisible	Is the tab currently visible?	
		VisitorKnown	Is the user known from former visits / are cookies enabled?	
		WindowHeight	The window height resolution of the client.	
		WindowWidth	The window width resolution of the client.	

Table A. Detailed overview of clickstream features used for each setting.

* Adopted from Buckinx and Van den Poel (2005) but adjusted to page categories prevalent in our own dataset.

** Adopted from Buckinx and Van den Poel (2005) but adjusted to a time horizon of the current session.

Table B reports the number of users predicted to leave without a purchase, which we assume will receive a coupon. This corresponds to a truncated confusion matrix for each of the 50 settings (2 shops x 5 feature sets x 5 conversion rates) depicted in Table 8. Given the profit setting and profit-based threshold selection, we observe that 1) the more accurate models based on the more sensitive feature sets tend to diverge more from the naïve benchmark, i.e. allow more coupons when the profitable baseline is to block coupons and vice versa, 2) optimization of the probability threshold interacts with the first effect to the extent that the total number of coupons may decrease if a profitable reduction of the costs of false positives can be achieved, for example for Shop 2 at 1% success rate when comparing *IntraSession* and *SessionBehavior* to the former additionally considering *CrossSession*. In line with the results concerning campaign revenue, building a model based on *IntraSession* features substantially increases the number of accurate classifications (compared to the no-coupon benchmark) or decreases the number of false classifications (compared to the all-coupon benchmark). Extending the information available to the model with more privacy-sensitive features further improves classification accuracy to a lesser extent.

Conversion rate		**1.0%**		**2.0%**		**3.0%**		**4.0%**		**5.0%**	
Model classification		**Accurate**	**False**	**Accurate**	**False**	**Accurate**	**False**	**Accurate**	**False**	**Accurate**	**False**
Shop 1	***Naïve: All/No Coupons***	0	0	0	0	0	0	194505	52820	194505	52820
	IntraSession	110651	6395	136288	10920	150685	15287	159870	18539	166162	21359
	& SessionBehavior	119707	7166	136856	10127	146198	12739	158171	17083	163026	19009
	& CrossSession	125022	6922	139719	9220	150451	11958	155632	13607	165473	17483
	& Identifiable	126189	7013	141992	9675	151172	12084	156570	13794	163558	16462
Shop 2	***Naïve: All/No Coupons***	0	0	0	0	202768	49018	202768	49018	202768	49018
	IntraSession	142002	13446	171859	18183	175798	19241	180311	20792	182813	21850
	& SessionBehavior	148622	14256	170202	17485	176862	19208	180095	20226	183113	21383
	& CrossSession	139185	10688	160643	14336	174033	17603	177612	18646	183899	21055
	& Identifiable	145473	11516	166243	15679	174108	17545	180030	19331	184462	21052

Table B. *Number of customers accurately and falsely by the respective model not to complete their purchase. Changes in the predictions between coupon success probabilities are due to profit-based probability threshold selection.*

In extension to Figure 2, we provide the correlation between the 25 features from any feature set with the highest variable importance in the random forest model (Figure A). The high correlation between the *SessionBehavior* and the *SessionContent* features related to shopping cart page visits stands out to explain the low marginal benefit from enriching the content data with behavioral data in our application. On the contrary, the *CrossSession* features are generally not correlated with other features depicted here and in particular not correlated with features from other feature sets.

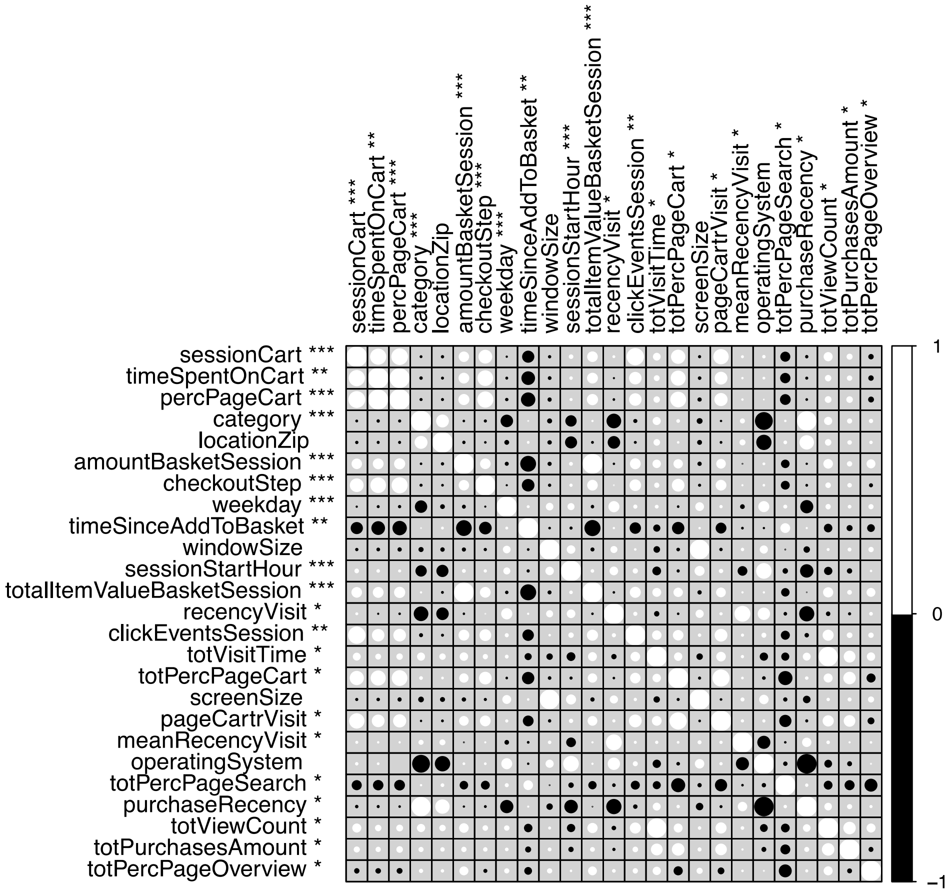

Figure A. *Correlation between the 25 features with the highest variable importance for the random forest model.*

References

Agrawal R, Srikant R (2000) Privacy-preserving data mining. ACM SIGMOD Record 29:439-450. doi: 10.1145/335191.335438

Akrivopoulou, C, Stylianou A (2009). Navigating in Internet: Privacy and the socioeconomic and legal implications of electronic intrusion. IGI Global.

Banerjee A, Ghosh J (2001) Clickstream clustering using weighted longest common subsequences. In: Proceedings of the Web Mining Workshop at the 1st SIAM Conference on Data Mining.

Bansal G, Zahedi F, Gefen D (2015) The role of privacy assurance mechanisms in building trust and the moderating role of privacy concern. European Journal of Information Systems 24:624-644. doi: 10.1057/ejis.2014.41

Baumer D, Earp J, Poindexter J (2004) Internet privacy law: A comparison between the United States and the European Union. Computers & Security 23:400-412. doi: 10.1016/j.cose.2003.11.001

Bennett PN, White RW, Chu W, Dumais ST, Bailey P, Borisyuk F, Cui X (2012) Modeling the impact of short-and long-term behavior on search personalization. In: Proceedings of the 35th international ACM SIGIR Conference on Research and Development in Information Retrieval, 185-194, ACM.

Boda K, Földes Á, Gulyás G, Imre S (2012). User tracking on the web via cross-browser fingerprinting. Information Security Technology for Applications, 31-46.

Breiman L (2001) Random forests. Machine Learning 45:5-32.

Buckinx W, Van den Poel D (2005) Predicting online-purchasing behaviour. European Journal of Operational Research 166:557-575. doi: 10.1016/j.ejor.2004.04.022

Chaffey D (2015) Digital business and e-commerce management, 6th ed. Pearson.

comScore (2007) Cookie-based counting overstates size of web site audiences. In: comScore, Inc. http://www.comscore.com/chi/Insights/Press-Releases/2007/04/comScore-Cookie-Deletion-Report. Accessed 22 Dec 2016

Cooley R, Mobasher B, Srivastava J (1997) Web mining: Information and pattern discovery on the world wide web. In: Proceedings of the Ninth IEEE International Conference on Tools with Artificial Intelligence, 558-567, IEEE.

Dinev T, Xu H, Smith JH, Hart P (2013). Information privacy and correlates: an empirical attempt to bridge and distinguish privacy-related concepts. European Journal of Information Systems, 22:295-316.

Eckersley P (2010) How unique is your web browser? In: International Symposium on Privacy Enhancing Technologies Symposium, 1-18, Springer Berlin Heidelberg.

Elkan C (2001) The foundations of cost-sensitive learning. In: International Joint Conference on Artificial Intelligence 17:973-978.

eMarketer (2016) Worldwide retail e-commerce sales will reach $1.915 trillion this year. In: Emarketer.com. https://www.emarketer.com/Article/Worldwide-Retail-Ecommerce-Sales-Will-Reach-1915-Trillion-This-Year/1014369. Accessed 22 Dec 2016

Gregorutti B, Michel B & Saint-Pierre P (2017) Correlation and variable importance in random forests. Statistical Computing 27:659-678. doi: 10.1007/s11222-016-9646-1

Greis F (2016) Browser-Addons: Browserverläufe von Millionen deutschen Nutzern verkauft. In: Golem.de. http://www.golem.de/news/browser-addons-browserverlaeufe-von-millionen-deutschen-nutzern-verkauft-1611-124171.html. Accessed 22 Dec 2016

Guo Q, Agichtein E (2010b) Ready to buy or just browsing? Detecting web searcher goals from interaction data. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 130-137, ACM.

Guo Q, Agichtein E (2010a). Towards predicting web searcher gaze position from mouse movements. In: Proceedings on Extended Abstracts on Human Factors in Computing Systems (CHI), 3601-3606.

Chan T, Joseph I, Macasaet C, Kang D, Hardy RM, Ruiz C, Porras R, Baron B, Qazi K, Hannon P, Honda T (2014) Predictive models for determining if and when to display online lead forms. In: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI), 2882-2889.

Hoofnagle C, Urban J, Li S (2012) Privacy and modern advertising: Most us internet users want ‘do not track’ to stop collection of data about their online activities. In: Amsterdam Privacy Conference.

Iwanaga J, Nishimura N, Sukegawa N, Takano Y (2016) Estimating product-choice probabilities from recency and frequency of page views. Knowledge-Based Systems 99:157-167. doi: 10.1016/j.knosys.2016.02.006

Jiang Q, Tan CH, Wei KK (2012) Cross-website navigation behavior and purchase commitment: A pluralistic field research. In: Proceedings of the Pacific Asia Conference on Information Systems (PACIS).

KantarMedia (2016) CPG digital coupon circulation grows by 23.4% in 1H16, reaching 3.7 billion. In: Kantarmedia.com. http://www.kantarmedia.com/us/newsroom/press-releases/cpg-digital-coupon-circulation-grows-by-23-4-in-1h16. Accessed 1 Mar 2017

Khajehzadeh S, Oppewal H, Tojib D (2014) Consumer responses to mobile coupons: The roles of shopping motivation and regulatory fit. Journal of Business Research 67:2447-2455. doi: 10.1016/j.jbusres.2014.02.012

Kim DJ, Ferrin DL, Rao HR (2008) A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. Decision Support Systems 44:544-564. doi: 10.1016/j.dss.2007.07.001

Kuhn M, Johnson K (2013). Applied predictive modeling. New York, Springer.

Lee M, Ferguson ME, Garrow LA, Post D (2010) The impact of leisure travelers' online search and purchase behaviors on promotion effectiveness. Working Paper, Georgia Institute of Technology.

Lessmann S, Voß S (2010) Customer-centric decision support: A benchmarking study of novel versus established classification models. Business & Information Systems Engineering 2:79-93. doi: 10.1007/s12599-010-0094-8

Libert T (2015) Privacy implications of health information seeking on the web. Communications of the ACM 58:68-77.

Lin E (2002) Prioritizing privacy: A constitutional response to the Internet. Berkeley Technology Law Journal 17:1085-1154.

Liu C, Marchewka J, Lu J, Yu C (2005) Beyond concern: a privacy–trust–behavioral intention model of electronic commerce. Information & Management 42:127-142. doi: 10.1016/j.im.2004.01.002

Lu L, Dunham M, Meng Y (2005) Mining significant usage patterns from clickstream data. In: Advances in Web Mining and Web Usage Analysis, 1-17, Springer Berlin Heidelberg.

Margineantu DD (2001) Methods for cost-sensitive learning. Doctoral dissertation, Department of Computer Science, Oregon State University.

Masand B., Piatetsky-Shapiro G (1996) A comparison of approaches for maximizing business payoff of prediction models. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining Portland, OR, USA, 195-201, AAAI Press Menlo Park.

Metzger M (2004) Privacy, trust, and disclosure: Exploring barriers to electronic commerce. Journal of Computer-Mediated Communication 9. doi: 10.1111/j.1083-6101.2004.tb00292.x

Moe W (2003) Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. Journal of Consumer Psychology 13:29-39. doi: 10.1207/153276603768344762

Moe W, Fader P (2004) Capturing evolving visit behavior in clickstream data. Journal of Interactive Marketing 18:5-19. doi: 10.1002/dir.10074

Moe WW, Chipman H, George EI, McCulloch RE (2002). A Bayesian treed model of online purchasing behavior using in-store navigational clickstream. Revising for 2nd review at Journal of Marketing Research.

Moertini VS, Ibrahim N (2015) Efficient techniques for predicting suppliers churn tendency in e-commerce based on website access data. Journal of Theoretical & Applied Information Technology, 74.

Montgomery A, Li S, Srinivasan K, Liechty J (2004) Modeling online browsing and path analysis using clickstream data. Marketing Science 23:579-595. doi: 10.1287/mksc.1040.0073

Nikiforakis N, Kapravelos A, Joosen W, Kruegel C. Piessens F, Vigna G (2014) On the workings and current practices of web-based device fingerprinting. IEEE Security & Privacy 12:28-36.

Nofer M, Hinz O, Muntermann J, Roßnagel H (2014) The economic impact of privacy violations and security breaches: a laboratory experiment. Business & Information Systems Engineering 6:339-348. doi:10.1007/s12599-014-0351-3

O'Connell BM, Walker KR (2014) User-browser interaction-based fraud detection system. In: USPTO Patent Full-Text and Image Database. http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.htm&r=1&f=G&l=50&s1=8,650,080.PN.&OS=PN/8,650,080&RS=PN/8,650,080. Accessed 22 Dec 2016

Padmanabhan B, Zheng Z, Kimbrough SO (2001) Personalization from incomplete data: what you don't know can hurt. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California.

Padmanabhan B, Zheng Z, Kimbrough SO (2006) An empirical analysis of the value of complete information for eCRM models. MIS Quarterly, 247-267.

Pai D, Sharang A, Yadagiri MM, Agrawal S (2014) Modelling visit similarity using click-stream data: A supervised approach. In: Web Information Systems Engineering (WISE), 135-145, Springer International Publishing

Park CH, Park YH (2015) Investigating purchase conversion by uncovering online visit patterns. SSRN Electronic Journal. doi: 10.2139/ssrn.1685469

Pitman A, Zanker M (2010). Insights from applying sequential pattern mining to e-commerce click stream data. In: IEEE International Conference on Data Mining Workshops (ICDMW), 967-975, IEEE.

Pollach I (2011) Online privacy as a corporate social responsibility: an empirical study. Business Ethics: A European Review 20:88-102.

Rodden K, Fu X, Aula A, Spiro I (2008) Eye-mouse coordination patterns on web search results pages. In: In Proceedings of Extended Abstracts on Human Factors in Computing Systems (CHI'08).

Sarwar SM, Hasan M, Ignatov DI (2015) Two-stage cascaded classifier for purchase prediction. arXiv preprint arXiv:1508.03856
Sato S, Asahi Y (2012) A daily-level purchasing model at an e-commerce site. International Journal of Electrical and Computer Engineering (IJECE). doi: 10.11591/ijece.v2i6.1816
Senécal S, Kalczynski P, Fredette M (2014) Dynamic identification of anonymous consumers' visit goals using clickstream. International Journal of Electronic Business 11:220. doi: 10.1504/ijeb.2014.063036
Senécal S, Kalczynski P, Nantel J (2005) Consumers' decision-making process and their online shopping behavior: a clickstream analysis. Journal of Business Research 58:1599-1608. doi: 10.1016/j.jbusres.2004.06.003
Sheng VS, Ling CX (2006) Thresholding for making classifiers cost-sensitive. In: Proceedings of the 21st National Conference on Artificial Intelligence. AAAI Press, Boston, MA, USA.
Sipior JC, Ward BT, Mendoza RA (2011) Online privacy concerns associated with cookies, flash cookies, and web beacons. Journal of Internet Commerce 10:1-16.
Sismeiro C, Bucklin R (2004) Modeling purchase behavior at an e-commerce web site: A task-completion approach. Journal of Marketing Research 41:306-323. doi: 10.1509/jmkr.41.3.306.35985
Skok G (2000) Establishing a legitimate expectation of privacy in clickstream data. Michigan Telecommunications and Technology Law Review 6:61-88.
Solove DJ (2001) Privacy and power: Computer databases and metaphors for information privacy. Stanford Law Review 53:1393-1462.
Stange M, Funk B (2014) Real-time-advertising. Business & Information Systems Engineering 6: 305-308. doi: 10.1007/s12599-014-0346-0
Stange M, Funk B (2015) How much tracking is necessary? – The learning curve in Bayesian user journey analysis. In: Proceedings of the 23rd European Conference on Information Systems.
Statista (2016a) Executive survey: big data sets that add the most value 2012 | Survey. In: Statista. https://www.statista.com/statistics/249054/executive-survey-on-big-data-sets-that-add-the-most-company-value/. Accessed 22 Dec 2016
Statista (2016b) Löschen oder Unterdrücken von Cookies bei deutschen Internetnutzern bis 2015 | Umfrage. In: Statista. https://de.statista.com/statistik/daten/studie/168870/umfrage/nutzung-von-programmen-die-cookies-loeschen/. Accessed 22 Dec 2016
Statista (2016c) Global online shopping conversion rate 2016 | Statistic. In: Statista. https://www.statista.com/statistics/439576/online-shopper-conversion-rate-worldwide/. Accessed 12 Jan 2017
Statista (2017) The ten coupon websites with the highest conversion rate in China in June 2011 | Statistic. In: Statista. https://www.statista.com/statistics/278752/coupon-websites-by-conversion-rate-in-china/. Accessed 08 Nov 2017
Suh E, Lim S, Hwang H, Kim S (2004) A prediction model for the purchase probability of anonymous customers to support real time web marketing: a case study. Expert Systems with Applications 27:245-255. doi: 10.1016/j.eswa.2004.01.008
Turow J, King J, Hoofnagle C, Bleakley A, Hennessy, M (2009) Americans reject tailored advertising and three activities that enable it. SSRN Electronic Journal. doi: 10.2139/ssrn.1478214
Van der Meer D, Dutta K, Datta A, Ramamritham K, Navanthe SB (2000). Enabling scalable online personalization on the web. In: Proceedings of the 2nd ACM Conference on Electronic Commerce, 185-196, ACM.
Vroomen B, Donkers B, Verhoef P, Franses P (2005) Selecting profitable customers for complex services on the Internet. Journal of Service Research 8:37-47. doi: 10.1177/1094670505276681
Wu F, Chiu IH, Lin JR (2005) Prediction of the intention of purchase of the user surfing on the web using hidden Markov model. In: Proceedings of International Conference on Services Systems and Services Management (ICSSSM'05). 1:387-390, IEEE.
Yang Y (2010) Web user behavioral profiling for user identification. Decision Support Systems 49:261-271. doi: 10.1016/j.dss.2010.03.001
Zhang Y, Bradlow E, Small D (2015) Predicting customer value using clumpiness: From RFM to RFMC. Marketing Science 34:195-208. doi: 10.1287/mksc.2014.0873
Zhao Y, Yao L, Zhang Y (2016) Purchase prediction using Tmall-specific features. Concurrency and Computation: Practice and Experience 28:3879-3894. doi: 10.1002/cpe.3720
Zheng Z, Padmanabhan B, Kimbrough S (2003) On the existence and significance of data preprocessing biases in web-usage mining. INFORMS Journal on Computing 15:148-170. doi: 10.1287/ijoc.15.2.148.14449

ARTICLE 10:

REVENUE UPLIFT MODELING

Authors: Robin Gubela, Humboldt-Universität zu Berlin, Berlin, Germany
Stefan Lessmann, Humboldt-Universität zu Berlin, Berlin, Germany
Johannes Haupt, Humboldt-Universität zu Berlin, Berlin, Germany
Annika Baumann, Humboldt-Universität zu Berlin, Berlin, Germany
Tillmann Radmer, Humboldt-Universität zu Berlin, Berlin, Germany
Fabian Gebert, Akanoo GmbH, Hamburg, Germany

Proceedings: 38th International Conference on Information Systems (ICIS'17)

Publication year: 2017

Abstract

The measurement of the effectiveness of a marketing campaign is a challenging task. Whereas established approaches do not consider causality, uplift models take into account which customers display some behavior because of the marketing action and model this target as differential response. The paper categorizes existing approaches toward uplift modeling collected from different fields into a conceptual taxonomy to establish the state-of-the-art and proposes a novel approach named revenue uplift modeling. Contrary to existing approaches, which model incremental response, revenue uplift models predict the incremental revenue with the goal to maximize the gain per marketing incentive for heterogeneous customers. An experiment based on a large real-world dataset of e-commerce shops across several industries provides a benchmark on the choice of machine learning methods to implement the identified uplift modeling approaches and demonstrates the effectiveness of the revenue uplift model in a real-world e-commerce environment.

Keywords: *Data Mining, Decision Analysis, Customer Targeting, Marketing Campaign Planning*

1. Introduction

Advertisements are omnipresent. A recent study of media use and advertisement exposure points out that the typical U.S. adult encounters a total of about 153 advertisements each day (Media Dynamics, Inc., 2014). Accordingly, advertising investment is substantial. In 2015 alone, about 161 billion USD were spent on digital advertising across all Internet-connected devices worldwide (eMarketer, 2016). To ensure accountability of investments and allocate marketing resources efficiently, it is important to measure the effectiveness of advertisement and more generally marketing communication. This remains a challenging undertaking. In particular, for a marketing stimulus to be judged effective, it should lead a customer to perform an intended action (e.g., purchase a product, download an app, sign-up for a newsletter, etc.). Data on individual customers, ad exposure, and customer conversion is often available, especially in online marketing. However, a co-occurrence of customer behavior and ad exposure is insufficient to conclude that the ad caused the observed customer action. Establishing such causal link represents a major obstacle in measuring marketing effectiveness (Rzepakowski and Jaroszewicz, 2012b).

A large body of literature examines data-driven models for customer targeting in offline settings, e.g. catalog marketing, and e-commerce, e.g. real-time couponing. Literature surveys in customer relationship management (CRM) (Ngai et al., 2009) and specific CRM tasks such as churn modeling (Verbeke et al., 2012) or direct marketing (Bose and Xi, 2009) illustrate the popularity of supervised machine learning methods to develop targeting models. Using data from a past campaign including explanatory variables (e.g., customer characteristics) and a response variable (e.g., whether a customer has churned or bought an item from a sales catalog), a learning method estimates a functional relationship between the response and explanatory variables. The estimated model facilitates predicting the value of the response from the explanatory variables (e.g., for novel customers). The uplift modeling community calls this approach response modeling because the model learns to recognize customers that have responded in the past (Radcliffe and Surry, 1999). Although widely used in the literature, the response modeling approach is flawed in that it disregards causality. For example, a customer may receive a special offer and buy the advertised product subsequently, but she may have bought the same product without the discount (Radcliffe, 2007). Uplift models overcome this inadequacy through predicting differential response; that is whether the customer buys because of the offer (Kane et al., 2014). Therefore, the uplift concept quantifies the true effectiveness of a campaign (Lo, 2002).

Uplift models support marketing managers in campaign planning and targeting marketing communication to customers who would not convert without the incentive (Soltys et al., 2015). This implies that an uplift model aims at estimating a causal link between a marketing action (e.g., offering a customer a special deal) and customer behavior (e.g., accepting the offer). Estimating the change in customer behavior that results from a solicitation, uplift models are especially suitable to support targeting decisions in campaign planning and increase campaign profitability (Siegel, 2011; Radcliffe and Surry, 2011; Larsen, 2010). An analysis of the literature on uplift models in marketing reveals that existing approaches focus on conversion and churn modeling, the goal of which is to win novel customers and prevent customer defection, respectively (Park and Park, 2016; Verbeke et al, 2012). In this regard, the strategic marketing objective behind current models is market share. In terms of the underlying uplift modeling methodology, conversion and retention models predict a dichotomous response variable using classification methods.

The paper extends previous literature through introducing revenue uplift modeling. A revenue uplift model predicts the incremental revenue that results from targeting a customer with a marketing message. In many applications, customers differ in their spending (Jacobs et al., 2016). Modeling revenue uplift accounts for this type of heterogeneity, which a conversion model is unable to accommodate. Therefore, revenue uplift modeling reflects the value-based idea of CRM (Reinartz and Kumar, 2003). Considering the focus of prior work on conversion uplift for customer acquisition and retention, revenue uplift modeling is also a relevant addition in that it provides an approach to target marketing campaigns that aim at increasing customer spending such as cross-/up-selling campaigns (Netessine et al., 2006).

In summary, the paper makes three contributions. First, existing approaches toward uplift modeling are categorized to sketch the field and highlight conceptual differences. This is useful since uplift modeling is still a niche topic in the academic literature. Second, a novel modeling strategy is proposed to predict revenue uplift. Targeting marketing communication so as to maximize revenue uplift is especially suitable for campaigns that aim at growing existing customers (e.g., cross-/up-selling). In that sense, the new approach naturally complements existing solutions for conversion and retention uplift modeling, which are geared toward customer acquisition and preventing customer defection, respectively. Third, a comprehensive empirical evaluation is carried out to demonstrate the effectiveness of the new uplift model in a real-world e-commerce environment. In addition to assessing alternative uplift modeling strategies, the experiment also provides original insights into the comparative performance of alternative machine learning methods for classification and regression to implement uplift models.

The results of the experiment confirm the effectiveness of the proposed approach. For the large e-commerce data set employed in the study, which comprises campaign results and actual sales from several e-shops across different industries, the new revenue uplift model provides the largest increase in incremental revenue and outperforms the benchmarks considered in the study. Although the model's uplift estimate is not unbiased, it's bias is somewhat lower compared to revenue models from challenger approaches because of the unique modification of the target variable. Furthermore, previous (conversion) uplift models are found ineffective in that they fail to outperform a simple response modeling approach. These results provide strong evidence that revenue uplift modeling is a useful technique to target marketing communication to responsive customers.

The paper is organized as follows: The next section introduces uplift modeling fundamentals before relevant prior work is revised. Subsequent sections elaborate on the proposed methodology and the experimental design. Afterwards, empirical results for conversion and revenue uplift modeling are reported, integrated, and discussed. The paper then concludes with a summary and outlook to future research.

2. Uplift Modeling Fundamentals and Process Model

The philosophy of an uplift-based targeting approach is that marketing communication should concentrate on customers who are influenced by the campaign (Rzepakowski and Jaroszewicz, 2012a). Rather than predicting customers' response probability and soliciting likely responders, as done in response/churn modeling (Chen et al., 2015; Neslin et al., 2006), the targeting decision should be based on the change in customers' likelihood to respond due to being targeted. These customers are called *Persuadables* in the literature and constitute the only group worth a marketing investment (Kane et al., 2014).

Identifying the treatment effect requires information on the response of individuals who have not received the treatment. Since each individual cannot be simultaneously treated and not-treated, the treatment effect is identified using the outcome observed in a control group. Therefore, an experimental setting with randomized treatment and control group is a prerequisite to develop an uplift model. This may be seen as a disadvantage compared to response modeling. However, in marketing and especially online marketing obtaining control group information is relatively straightforward. In particular, A/B testing is a popular approach to perform random experiments in e-commerce. For example, a website owner may randomly assign visitors to different groups each of which get to see a different version of the homepage, hoping that the random assignment facilitates causal statements as to the effectiveness of the different page versions.

To the best of our knowledge, the literature on uplift models relies exclusively on this approach of treatment-control group comparisons to establish causality. However, it is important to note that A/B tests may fail to implement a statistically sound random experiment, especially in high-dimensional settings, which may invalidate conclusions on causal links, and may be impractical in large-scale settings where a vast number of tests are performed in parallel (Kohavi et al., 2013). For consistency with previous literature on uplift modeling, we focus on randomized trials in the form of A/B tests as vehicle to establish causal relationships. Evaluating other causal inference procedures such as, for example propensity scores or instrumental variables (Imbens, 2004) for uplift modeling is a fruitful area of future research but beyond the scope of this paper.

A/B tests are used to estimate the marginal performance increase due to a marketing incentive, the uplift, but also facilitate the training of models based on this metric (e.g. Radcliffe and Surry, 2011). In Figure 1, we summarize the concept of uplift with the four-fields target matrix from Kane et al. (2014). Response models distinguish between responders and non-responders (left and right column) irrespective of the actual effect of treatment. The goal of uplift models is to use the information on the control population to also account for variation in response rate dependent on whether the marketing incentive was received. In other words, uplift models identify likely treatment responders (upper right), who respond specifically due to the marketing incentive and would not respond otherwise.

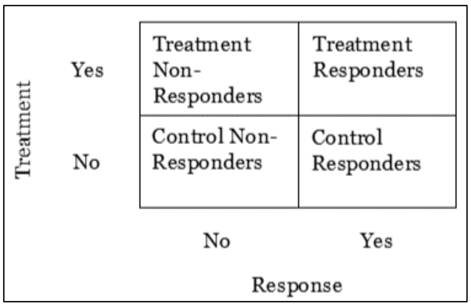

Figure 1. The four-fold target matrix

To formalize the methodological difference between uplift and response modeling, let $\boldsymbol{X_i} = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a vector of characteristics (i.e., explanatory variables) of customer i, and let $Y_i \in \{0,1\}$ be a binary response variable, for example whether customer i bought a product in a previous campaign. Uplift models build on the concept of A/B testing, meaning that customers are divided into two groups: treatment and control (Kohavi et al., 2009). Let $T_i \in \{0,1\}$ be an indicator variable of the group membership of customer *i,* with $T_i = 0$ and $T_i = 1$ indicating membership to the control and treatment group, respectively. Then, with $P(Y_i|\boldsymbol{X_i}, T_i = 1)$ and $P(Y_i|\boldsymbol{X_i}, T_i = 0)$ denoting customer-level probabilities in the corresponding groups, traditional response models predict the conditional probability $P(Y_i|\boldsymbol{X_i}, T_i = 1)$, whereas an uplift model predicts the change in behavior resulting from a treatment $P(Y_i|\boldsymbol{X_i}, T_i = 1) - P(Y_i|\boldsymbol{X_i}, T_i = 0)$. In marketing, the treatment can be an advertisement, direct mail, or some other marketing action. Many supervised learning methods are available to estimate conditional response $P(Y_i|\boldsymbol{X_i})$ (Hastie et al., 2009).

An intuitive approach to develop an uplift model involves estimating two models to predict $P(Y_i|\boldsymbol{X_i}, T_i = 1)$ and $P(Y_i|\boldsymbol{X_i}, T_i = 0)$, respectively. Campaign planners can then calculate the uplift for individual customers as the difference between these models' predictions and target customers in the order of their estimated uplift. This approach is known as the two-model or indirect approach (e.g. Lo and Pachamanova, 2015). The indirect approach embodies the objective to maximize responders in the treatment group while minimizing control group responders but suffers important limitations. First, estimating two models increases computational costs. Second and more importantly, the distribution of the difference of the probabilities is often different from the distribution of the respective probabilities, which causes bias and poor model performance (Chickering and Heckerman, 2000; Rzepakowski and Jaroszewicz, 2012b).

The shortcomings of the indirect approach led to the development of improved uplift modeling regimes. Furthermore, the distinction of treatment and control group customers has implications for all stages of the model development process. To systematize related work in the field and identify the contribution of the paper, the uplift modeling process for marketing (UMPM) is introduced in Figure 2. The process model is based on the well-known KDD process (Fayyad et al., 1996).

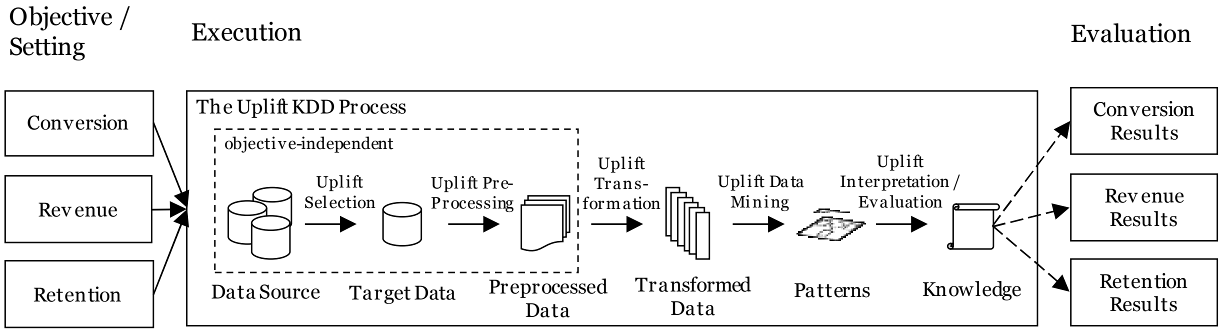

Figure 2. Uplift modeling process for marketing

Prior work on uplift modeling in marketing focuses predominantly on conversion uplift. Only two studies examine retention uplift (Guelman, 2014; Siegel, 2013). Revenue uplift modeling has not been considered at all but is introduced here. The UMPM strives to raise the awareness of different modeling objectives in campaign planning. To that end, the UMPM distinguishes three stages: (1) selecting a suitable business objective out of conversion, revenue, or retention modeling for a specific campaign, (2) pursuing the chosen objective to gain insight (along the stages of the UMPM), and finally (3) evaluating results in recognition of the campaign objective to identify and target the truly responsive customers with the next marketing campaign.

All campaign goals in Figure 2 imply a profit objective. Winning new customers with conversion modeling increases revenues, even if the magnitude of the increase is not the focus of attention. Cross-/up-selling campaigns and campaigns aiming at customer growth in general maximize revenue directly, while campaigns to prevent customer attrition sustain future revenues. Clearly, none of the objectives and underlying uplift modeling strategies is generally preferable. Rather, the point of the UMPM is to stress that campaign planners who use uplift models to support targeting decisions should choose a definition of uplift that best matches the campaign objective and then develop a corresponding model. For example, when customer spending varies substantially and there is a small fraction of high value customers, a revenue uplift model will recommend a smaller campaign size than a conversion uplift model, which maximizes incremental sales. The smaller and more focused campaign is likely to be more profitable because it avoids the costs of soliciting low-value customers. This view is supported by the empirical results of this study.

The UMPM has been designed to provide maximal flexibility in the choice of the objective based on the respective specific marketing situations of campaign planners. Therefore, while the goal of the next campaign could be conversion-/retention-related, a rather value-related aim could be operationalized in another post-initiative (or vice versa). This single-campaign focus is typically not supported by other revenue-based models such as the customer lifetime value (CLV) which pre-empt decisions due to long-term strategies. Furthermore, CLV models are typically considered if long-term contractual agreements result from the desired action, which is rather the case for insurance or banking products/services than for fast moving consumer goods in e-retail. This is not least the case because of the accumulated value a customer generates if being locked-into a long-run agreement. In contrast, the buyer-seller relationship in e-commerce is typically rather transactional, which is why CLV models are rarely applied in this field. One might also argue that the focus on long-run customer relationships, as embodied in CLV models, is more geared toward tactic/strategic marketing management, whereas uplift models with their short-term campaign planning objectives (see Figure 2) are a tool for operational marketing planning. For example, measuring the causal influence of a marketing action on customer-level CLV is a complex undertaking., because changes in long-term strategic performance indicators like customer-level CLV and customer equity, respectively, can only be observed in the longer run where a multitude of external factors will simultaneously affect these indicators, leading to serious modeling issues with respect to endogeneity.

Figure 2 indicates that the selection of a campaign objective has methodological implications. Multiple stages in the model development process depend on the objective. Most importantly, the response variable Y_i is dichotomous in conversion and retention modeling (success/failure to convert/retain customer) and continuous in revenue modeling (purchase amount). Accordingly, conversion/retention uplift models require classification methods to estimate conditional response $P(Y_i|\boldsymbol{X_i})$ whereas revenue uplift models use regression methods (Hastie et al., 2009). Subsequent parts of the paper will further detail objective-specific modeling implications.

3. Related Literature

The review of prior work is organized along the stages of the UMPM (Figure 2). In general, specific modeling challenges arise in uplift modeling due to the estimation of causal effects. For example, the distinction of customers into treatment and control group affects data selection (Kane et al., 2014) as well as preparatory activities including the handling of missing values, outliers and feature selection (Hua, 2016; Yong, 2015; Larsen, 2010; Hansen and Bowers, 2008). It also affects the evaluation of uplift models, which often grounds on a comparison between model predictions for treatment and control group customers (Nassif et al., 2013; Radcliffe and Surry, 2011; Radcliffe, 2007). Data transformation is important for uplift modeling because a suitable transformation of the explanatory variables or the response facilitates predicting uplift using standard learning methods (e.g. Tian et al., 2014; Lo, 2002).

An alternative strategy is to modify existing learning methods. In the spirit of the KDD process, an algorithmic modification exemplifies uplift data mining, which represents the prevailing approach in prior work. Corresponding studies strive to estimate uplift directly using tree-based algorithms with adapted splitting and pruning criteria (e.g. Hansotia and Rukstales, 2002), ensembles of uplift decision trees (Guelman et al., 2015), artificial neural networks (Manahan, 2005), k-nearest neighbours (Larsen, 2010) and support vector machines (Jaroszewicz and Zaniewicz, 2016; Zaniewicz and Jaroszewicz, 2013).

The paper focuses on uplift transformation. Compared to uplift data mining, approaches for response and covariate transformation are generic. As will be detailed below, they facilitate an implementation of the modeling methodology using conventional machine learning methods. Given that this is the first paper to study revenue uplift modeling, it is useful to compare a broad set of different regression methods. Such comparison can identify methods that work well for revenue uplift. Future work could then develop modification of these methods to approach the revenue uplift modeling problem directly. In contrast, it seems less suitable to start the journey into revenue uplift modeling with a modification of one regression method, arbitrarily chosen from a vast space of alternative methods (Hastie et al., 2009).

4. Literature

4.1. Uplift Taxonomy

The uplift transformation framework (Figure 3) formally introduces and contextualizes revenue uplift modeling in the data transformation stage of the UMPM. The tree provides marketing analysts two options, a transformation of the input space (i.e., covariates) or the output space (i.e., the response variable). Response transformation can be further distinguished in terms of the underlying modeling objective.

If the objective is to increase conversion or retention rates, the response is a binary indicator variable which equals one if a customer has shown the focal behavior (has converted/churned) and zero otherwise. Response models rely exclusively on this information. Uplift models for conversion also predict a binary response variable but alter the group definition to model incremental conversions. The underlying learning methods are the same as those used in response modeling (e.g., logistic regression, neural networks, etc.). The two main transformation approaches are the class variable transformation (CVT) (Jaskowski and Jaroszewicz, 2012) and Lai's weighted uplift method (LWUM) (Lai, 2006). The paper focuses on the latter approach because recent benchmarking results indicate that it often outperforms alternative techniques (Kane et al., 2014).

Targeting models for revenue uplift transforms an originally continuous response variable (here, the revenue per customer) using information on whether customers were part of the treatment or control group. Depending on the specific transformation strategy, the new response can be continuous or binary. Drawing inspiration from previous work concerning the advantages of classification over regression models in direct marketing (Bodapati and Gupta, 2004), the proposed response discretization approach (RDT) produces a binary modeling target. However, an intermediate step in the novel approach, which grounds on Jaroszewicz (2016), delivers a continuous transformed response variable, which offers an alternative route to develop a revenue uplift model. A methodological difference between the two approaches is that RDT works with classification methods whereas the alternative relies on regression methods.

The literature proposes two approaches for covariate transformation; the interaction term method (ITM) (Lo, 2002) and the treatment-covariates interactions approach (TCIA) (Tian et al., 2014). Conceptually, both approaches are similar and differ only in the scaling of the response and normalization of the explanatory variables. In view of this, the empirical analysis includes the more recent TCIA approach.

Note that covariate transformation can be combined with response transformation. Thus, there are four options to build uplift models using covariate transformation. Either models are built on the untransformed conversion variable (conversion response modeling with modified covariates), the untransformed revenue variable (revenue response modeling with modified covariates), the transformed conversion variable (conversion uplift modeling with modified covariates) or the transformed revenue variable (revenue uplift modeling with modified covariates). The two latter options are illustrated with the dotted arrow between the covariates transformation and response transformation boxes in Figure 3.

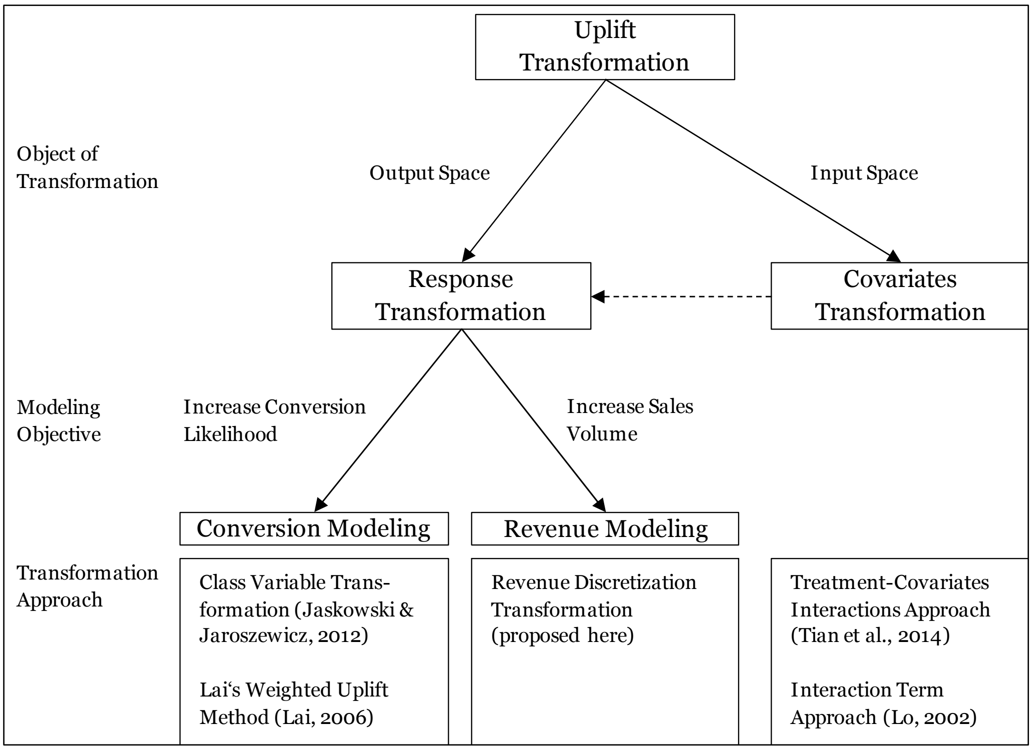

Figure 3. Uplift transformation framework.

4.2. Underlying Approaches Detailed

4.2.1. Conversion Response Transformation

The LWUM approach transforms the response variable so as to facilitate the use of conventional classification models to predict conversion uplift. Let $z_{\mathrm{i},c}$ be the binary transformed response of customer i, with c identifying the campaign objective (i.e., conversion). The response $z_{\mathrm{i},c}$ equals one for treatment group customers who convert and control group customers who do not convert. Both states represent a success (Lai, 2006). In all other cases, $z_{\mathrm{i},c}$ is set to zero. Formally, this logic is captured in:

$$\mathrm{z_{i,c}} = \begin{cases} 1 & \text{if } \mathrm{T_i} = 1 \cap \mathrm{Y_{i,c}} = 1 \cup \mathrm{T_i} = 0 \cap \mathrm{Y_{i,c}} = 0 \\ 0 & \text{otherwise} \end{cases}$$

with the transformed response $z_{\mathrm{i},c} \in \{0,1\}$. Recall that $T_i \in \{0,1\}$ is an indicator variable for control/treatment group, $Y_{i,c} \in \{0,1\}$ the original response variable, which captures the status of customer i (no conversion/conversion), and $i = 1, \ldots, N$ indexed customers in a past campaign of size N and $\boldsymbol{X_i} = (X_1, \ldots, X_n) \in \mathbb{R}^n$ is a vector of covariates. With LWUM, uplift is defined as:

$$Uplift_i^{Lai} = P\left(z_{i,c} = 1 \middle| \boldsymbol{X_i}\right) \cdot \mathrm{w_{pos}} - P\left(z_{i,c} = 0 \middle| \boldsymbol{X_i}\right) \cdot w_{neg}$$

where $\mathrm{w_{pos}}$ and $\mathrm{w_{neg}}$ are weighting parameters determined by the ratio of positive or negative cases in the data, respectively. For $\mathrm{w_{pos}} = w_{neg} = 1$, this approach reduces to the CVT introduced by Jaskowski and Jaroszewicz (2012).

4.2.2. Revenue Response Transformation

From an analytical point of view, the key feature that distinguishes conversion and revenue uplift is the target variable: Instead of transforming the (binary) conversion variable $Y_{i,c}$, the (continuous) revenue variable $Y_{i,r}$ is subject to transformation. In particular, let $Y_{i,r} \in \mathbb{R}$ be the original response revenue variable capturing sales revenue of customer *i,* with *r* once again indicating the primary objective of a campaign.

The proposed RDT approach for revenue uplift modeling is based on the concept to discretize a continuous response in order to decrease the bias due to incorrect model specification and increase prediction accuracy (Bodapati and Gupta, 2004). Although the authors consider a response modeling setting, their finding appears relevant for uplift modeling as well. When correctness of a model's specification cannot be ensured, which is often the case in real-world data due to factors such as omitted variables, the resulting bias in OLS estimation can be reduced through a discretization of the target variable at the expense of an increase in variance. In large sample sizes, the importance of variance, however, diminishes. Bodapati and Gupta (2004) gain this insight in simulation experiments with a maximum of 20,000 observations. Much larger sample sizes occur when targeting marketing communication in online environments and/or running campaigns to increase sales in e-commerce.

The logic behind value discretization can be illustrated with the sales situation of a book club (Bodapati and Gupta, 2004). Instead of predicting the annual number of books for all customers individually, the managerial challenge is to predict whether this number exceeds a pre-defined threshold. The actual task is then to determine the value of a discretizing function, $d(y)$, which the authors define as:

$$\mathrm{d(y)} = \begin{cases} 0 & \text{if } \mathrm{y} \in (0, \mathrm{y_{threshold}}] \\ 1 & \text{if } \mathrm{y} \in (\mathrm{y_{threshold}}, \infty) \end{cases}$$

with $y_{threshold}$ as the set value of the absolute number of books in the example. Supervised classification models facilitate estimation of this function (Bodapati and Gupta, 2004).

The RDT approach proposed in this paper combines the idea of revenue uplift modeling with the target design from conversion modeling in a multi-layer transformation scheme. The revenue variable $Y_{i,r}$is first transformed to obtain $z_{\mathrm{i},r}$ (Jaroszewicz, 2016) and then this variable is discretized to receive $z_{\mathrm{i,rg}}$. More formally, the two-step transformation corresponds to:

$$\mathrm{z_{i,r}} = \begin{cases} +\mathrm{Y_{i,r}} & \text{if } \mathrm{T_i} = 1 \cap \mathrm{Y_{i,r}} > 0 \cup \mathrm{T_i} = 1 \cap \mathrm{Y_{i,c}} = 1 \\ -\mathrm{Y_{i,r}} & \text{if } \mathrm{T_i} = 0 \cap \mathrm{Y_{i,r}} > 0 \cup \mathrm{T_i} = 0 \cap \mathrm{Y_{i,c}} = 1 \\ 0 & \text{otherwise.} \end{cases}$$

with $z_{\mathrm{i},r} \in \mathbb{R}$ as the transformed revenue that captures additional information from the group membership indicator. In particular, $z_{\mathrm{i},r}$ is equal to the original sales revenue for treatment group customers who made a purchase, equal to the negative sales revenue for control group customers who made a purchase, and zero otherwise. This transformation produces a novel response variable for direct uplift modeling. A single regression model suffices to predict $z_{\mathrm{i},r}$ which itself possesses all necessary information for uplift predictions. For RDT, however, $z_{\mathrm{i},r}$ is only an intermediate step. Rather than predicting $z_{\mathrm{i},r}$ with regression methods, a discretization procedure on $z_{\mathrm{i},r}$ facilitates use of classification methods and, more importantly, has the option to capitalize on the advantages of value discretization (Bodapati and Gupta, 2004).

$$\mathrm{z_{i,rg}} = \begin{cases} 0 & \text{if } \mathrm{z_r} \in (-\infty, 0] \\ 1 & \text{if } \mathrm{z_r} \in (0, \infty) \end{cases}$$

where $z_{\mathrm{i,rg}} \in \{0,1\}$. The key differentiating factors to Bodapati and Gupta's (2004) discretization are that the response variable has been pre-transformed and that negative numbers are captured in $z_{\mathrm{i,rg}}$, because customers who converted without having received a certain treatment are included. This points out that in $z_{\mathrm{i,rg}}$ information related to the treatment and control group is provided which underlines its characteristic of reflecting change in behavior because of having received a treatment.

The reason why the threshold has been set to zero is related to the objective in the context of uplift modeling. A “failure” is defined by $z_{\mathrm{i,rg}} = 0$. Customers who display the behavior intended by the marketer but without having received the treatment ($z_{\mathrm{i},r} = -Y_{i,r}$) and customers from both treatment and control group with zero purchases ($z_{\mathrm{i},r} = 0$) belong to this category. In contrast, “success” is related to customers who have purchased a product with the causal connection to the campaign treatment ($z_{\mathrm{i},r} = +Y_{i,r}$). This group is the only one that fulfills the condition $0 < z_{\mathrm{i},r} < \infty$ since the price of a product always starts at one cent and is never infinite. Compared to other approaches such as CVT and LWUM, RDT defines “success” differently in terms of the four-fields target matrix presented in Figure 1. In this regard, the only group to target depicts the treatment responders and not, in addition, control non-responders.

4.2.3. Covariates Transformation

Covariates transformation deals with the transformation of the input space. In case of TCIA, a dummy variable $T_i^* \in \{-1; +1\}$ is created. Its value depends on whether the customer has been in the treatment or control group. Then, T_i^* is multiplied with each of the n covariates to determine the interaction term, i.e. $T_i^* * X_i^*$ where X_i^* is modified using a mean centering procedure (Long, 1994). This additional term is taken into account when building uplift models. Following the idea by Lo (2002), the general design to model uplift is $E(Y_i|X_i) = f(T_i, X_\mathrm{i}, T_i * X_i)$ which can be further substituted into $E(Y_i|X_i, T_i = 1) - E(Y_i|X_i, T_i = 0)$. In the next step, TCIA takes each element from the input space and transforms it using $Z_i = T_i^* X_i^*/2$ which is used to predict the response based upon the modified covariates Z_i.

5. Experimental Design

5.1. Data and Experimental Setting

The experimental setting is based on a real-time targeting process in e-commerce. When customers visit the website of an e-commerce shop, a subset of selected customers receive an e-coupon at some point during their session with a discount of 10% off the final basket value. Each targeted customer receives a unique coupon code which needs to be used during the check-out process in the basket to activate the discount. Coupons are commonly used in digital marketing to simulate conversion and generate additional sales (e.g., Khajehzadeh et al. 2014).

While clicking through the website, a visitor is randomly assigned to either being scored by a random process or by a model, i.e. all customers are subject to pre-screening determining if they are eligible for the coupon campaign. Only those customers are further considered who have a high likelihood of responding to the coupon. In the next step, customers having a high likelihood of responding are randomly assigned to the treatment or control group. Customers in the treatment group receive a coupon, those in the control group do not receive a coupon. This process provides the treatment and control setting required for uplift modelling (Figure 4). The filtering stage, which identifies customers with a high likelihood of response, creates a selection bias towards more likely purchasers in the overall group resulting in a quasi-experiment.

A partner from industry provided the real-world data which is based on twenty-five different e- shops. There are 3,051,990 observations per variable and 62 variables. Each observation represents an individual customer session. The variables mainly capture customer-specific information such as key areas of the websites visited, including related length of time information. The data also covers the group membership indicator, shop-ID, time stamp together with information on (raw) conversions and basket values.

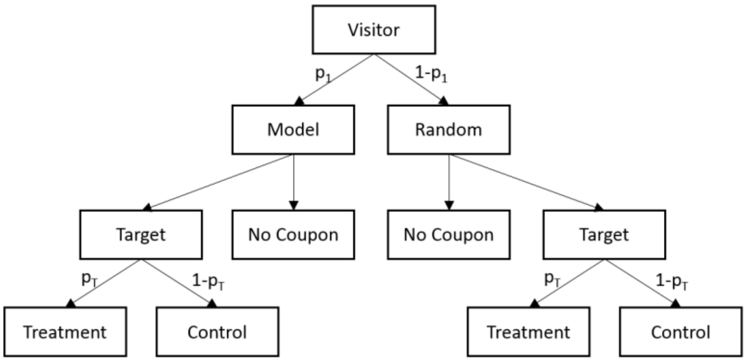

Figure 4. Treatment / control group assignment process for the dataset

Table 1 summarizes (i) the fraction of visitors in the treatment and control group, (ii) how many visitors of each group have made a purchase, and (iii) the overall uplift on the dataset based on the group differences in conversion rates. From the table's last column, it can be concluded that the overall uplift on the real-world dataset for the experiments is low. This suggests that the specific coupon offer is not particularly effective in increasing conversion behavior. However, this does not affect the suitability of the data since the focus of the paper is on uplift modeling strategies and thus the relative gain in conversions/revenues due to an improved targeting strategy.

Since the primary objective of the paper is to introduce revenue uplift modeling and the novel RDT approach in particular, it is interesting to examine the consequences of the steps in RDT on the revenue distribution. This analysis is shown in Figure 5. Besides the group membership distribution (left panel), the distribution of the baseline revenue response Y_r is presented (upper right). Moreover, the two smaller plots on the bottom highlight the distributions of the transformed revenue response without discretization (z_r) and after discretization (z_{rg}), respectively. Note that the analysis is based on a sub-sample of the whole data set (~ 420,000 observations) which is representative to the data used in the empirical study of this paper.

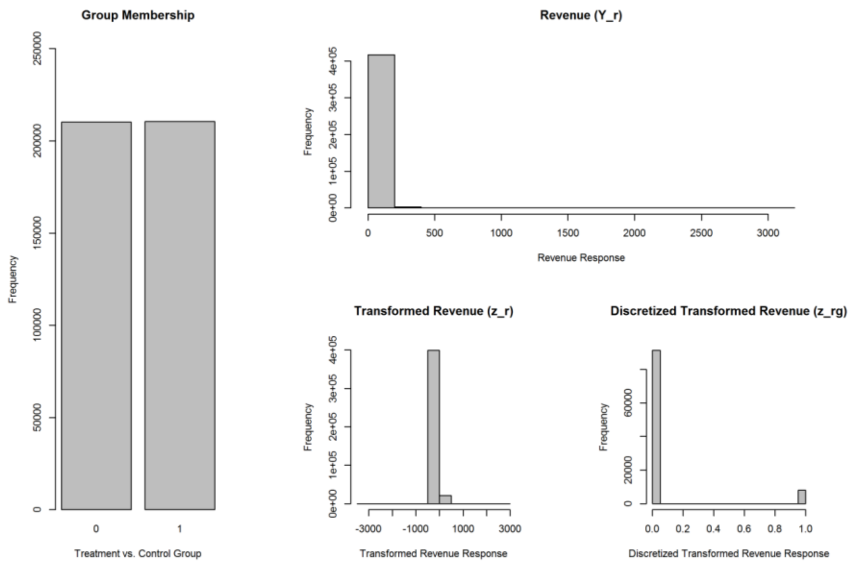

Figure 5. Implications on distributions of revenue response transformations.

The horizontal axes for both the Y_r and z_r plots extend to show the most extreme observed values, even if their frequency is too low to be displayed without scaling. Thus, for Y_r the minimum for revenue is 0€ and its maximum is around 3,000€ and for z_r the minimal value is -3,000€ and the maximal value 3,000€ (due to the transformation logic of the first step of RDT). It is noteworthy that there are few cases where a treatment led to a purchase of a high-priced product. The same is true for control group customers who have purchased a product starting in the price category of 1,000€ (i.e., generating high revenue hereby). This is because the high values (with and without inverted sign) occur so rarely. Since the comparably most frequent values of z_r are either negative or zero, the discretized transformed response z_{rg} is mainly zero. Only in few cases, i.e. when $z_r \in (0, \infty)$, it holds that $z_{\mathrm{rg}} = 1$. This is visualized in the bottom right chart.

Group	**Share on Dataset**	**Number of Observations**	**Number of Converters**	**Conversion Rate**	**Uplift**
Treatment	74.9%	2,285,835	175,791	7.69%	0.22%
Control	25.1%	766,155	57,285	7.47%	
Total	100%	3,051,990	233,076		

Table 1. Uplift on dataset.

5.2. Base Learners

Alternative uplift modeling approaches are implemented using supervised learning methods. Table 2 lists the methods that have been considered in the experiments. The selection of methods includes well-established individual learners (e.g. logistic regression and tree-based learners) and ensemble algorithms (e.g. random forest and gradient boosting). Interested readers find a comprehensive description of these methods in Hastie et al. (2009). In addition, Table 2 includes some methods that have recently shown promising results, especially in medical and biological informatics research (e.g. Extremely Randomized Trees, compare Nattee et al., 2017; Soltaninejad et al., 2016; Gotz et al., 2014) or seem to be often overlooked despite their advantages (e.g. Theil-Sen Regression, see Fernandes and Leblanc, 2005).

Many learning methods exhibit meta-parameters to adapt an algorithm to a particular data set (Hastie et al., 2009). Such parameters are tuned using grid-search, for which candidate parameter values have been obtained from literature (Baumann et al., 2015).

5.3. Validation Strategy

The whole dataset has been partitioned into a training set (~40%), a meta-parameter optimization set (~30%) and a validation set (~30%). In a first step, the models of all approaches are built on the training set and tested on the parameter optimization set to identify the optimal parameter configuration for the respective models. In a second step, the best models are trained on the training and parameter optimization set together (covering 70% of the whole dataset) and tested on the validation sample.

5.4. Performance Measures

Measures to assess predictive models are based on a comparison of actual and predicted outcomes for every individual unit of observation (Hastie et al., 2009). In uplift modeling, however, such comparison is impossible since no customer can receive and not receive a treatment at the same time (Radcliffe and Surry, 2011; Radcliffe, 2007). This phenomenon is known as the fundamental problem of causal inference (Holland, 1986). To evaluate uplift models, Qini curves and the corresponding Qini values have been developed. They can be considered an extension of cumulative gain charts and the corresponding Gini coefficient, which facilitate an assessment of response models (Radcliffe, 2007). Gain analysis assesses models in terms of cumulative increase of responses that follow from a model-based compared to a random targeting. For the standard lift metric, *gain* is defined as the number of conversions or the value of these conversions for response and revenue models, respectively, while the uplift metric considers the *incremental* or relative gain as compared to the control group.

The performance of uplift models is visualized using Qini curves by plotting the incremental gain against the percentage of the population that is targeted. Incremental gain is determined by, first, ordering the population by their model score and segmenting customers into groups with decreasing predicted response probability. Second, the incremental gain within each segment is calculated as the difference between responders (or revenue) in the treatment group and control group adjusted for the size of the groups.

The Qini coefficient provides a single number of model performance, which is useful to compare alternative models. To calculate the Qini coefficient, the Qini curve of a model is compared to a random model (Radcliffe and Surry, 2011). The performance line of the latter starts in the coordinate system's origin and ends up in (N, n) with N as the population size and n as the total incremental number of purchases (conversion modeling) or total incremental revenue (revenue modeling) if everyone is targeted instead of a certain subpopulation (Radcliffe, 2007). The random model poses a useful baseline that relevant models need to outperform to generate value. The Qini values Q is defined as the area between the model gain curve and the random model (diagonal line). It can be understood as an absolute measure of incremental gain. For clarity, we denote the Qini values for the two modeling objectives, i.e. incremental number of purchases for conversion modeling and incremental revenue for revenue modeling, by Q_c and Q_r respectively.

Conversion Modeling	**Revenue Modeling**
Logistic Regression (LogR) Calibrated Linear Support Vector Machine (SVM) k-Nearest-Neighbors (KNN) Naïve Bayes (NB) Stochastic Gradient Descent for Classification (SGDC) Random Forest for Classification (RFC) Calibrated Random Forest for Classification (RFC-C) Extremely Randomized Trees (ERT) Gradient Boosting for Classification (GBC)	Linear Regression (LinR) Ridge Regression (Ridge) Lasso Lars Regression (LL) Stochastic Gradient Descent for Regression (SGDR) Theil-Sen Regression (TS) Random Forest for Regression (RFR) Extremely Randomized Trees (ERT)

Table 2. Base learners.

A limitation of Q may be seen in the fact that different parts of gain/Qini curve carry different relevance to marketing practice. Campaigns are typically target to a small fraction of customers. Thus, the gain of a model for smaller targeting fractions is particularly important. Ling and Li (1998) proposed a weighting procedure to account for this issue in response modeling. We adopt their approach for uplift modeling. In formal terms, let Q_{wc} and Q_{wr} be the weighted scores across deciles of a certain model for conversion and revenue modeling, respectively.

Then, $\mathrm{Q_{wc}} = \frac{(0.9*\mathrm{Q_{1,c}}+0.8*\mathrm{Q_{2,c}}+\cdots+0.1*\mathrm{Q_{9,c}})}{\sum_{\mathrm{i}} \mathrm{Q_{i,c}}}$ and $\mathrm{Q_{wr}} = \frac{(0.9*\mathrm{Q_{1,r}}+0.8*\mathrm{Q_{2,r}}+\cdots+0.1*\mathrm{Q_{9,r}})}{\sum_{\mathrm{i}} \mathrm{Q_{i,r}}}$ with c and r indicating conversion and revenue, respectively, and $i = (0, 1, \ldots, 9)$ representing a decile index.

The following chapters present the experiments using the above performance measures. These are (i) Qini curves and Qini values Q_c and Q_r, (ii) their weighted versions Q_{wc} and Q_{wr} and (iii) incremental revenue.

6. Conversion Modeling

In terms of conversion modeling, we consider LWUM, TCIA and response modeling. Response modeling serves as benchmark that disregards the uplift philosophy. In total, 316 different classification models have been developed per approach, using the learning methods outlined above. Accordingly, a total of 948 classifiers are compared in the experiment. Some models have returned comparably biased probabilities. To address this problem, probability calibration based on Platt Scaling (Platt, 1999) and isotonic regression have been used for certain linear support vector machines and random forests, respectively. Figure 6 depicts model performance in terms of Qini curves per uplift modeling approach. The legend in each plot also provides Qini values.

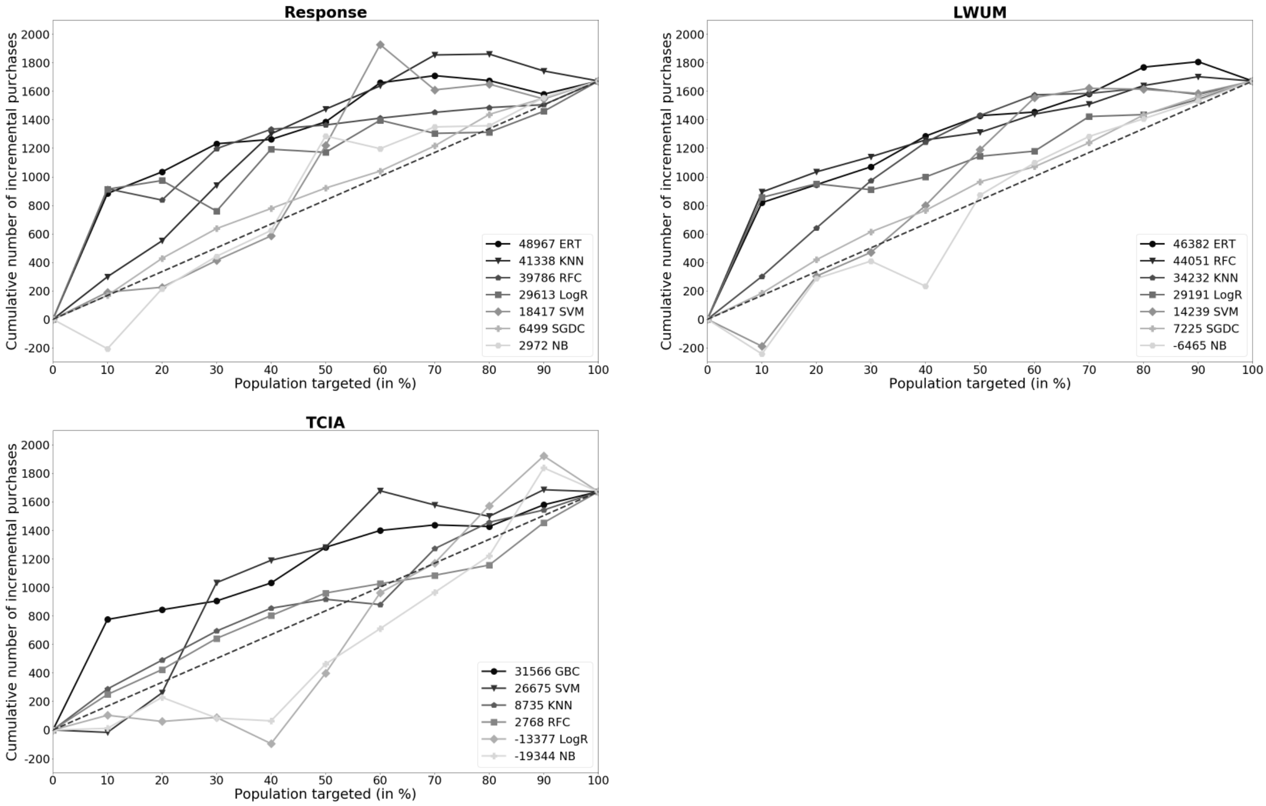

Figure 6. Best models per approaches for conversion modeling.

Figure 6 indicates that most of the uplift models succeed in outperforming the naïve benchmark, which is represented by the diagonal line. However, uplift models built on top of a Naïve Bayes classifier deviate from this pattern and typically perform weaker than the naïve benchmark. In this sense, Figure 6 warrants the conclusion that Naïve Bayes is not a suitable approach for this type of learning problem and should be avoided. Although its performance is better than that of Naïve Bayes, stochastic gradient descent for classification appears to be another candidate learner which proofs inadequate for the focal prediction task. The corresponding Qini curve falls sometimes below the naïve benchmark and never exceeds it with substantial margin. On the other hand, tree-based ensemble classifiers are among the best classifiers and show consistently good results across all uplift modeling approaches. The same applies to the KNN classifier, which is always among the top three methods per approach. This result is surprising in that KNN is a rather simple classifier.

A positive result shown in Figure 6 is that several of the considered uplift models display a steep increase in performance within the first decile. It is common practice in marketing to target only a small subset of the customer base with a campaign. Therefore, the degree to which a model delivers high uplift in the first decile (i.e., succeeds in identifying a small subset of highly responsive customers) is of paramount importance for campaign planning practice.

To simplify comparisons of alternative uplift modeling approaches to each other, Table 3 reports the per-decile-uplift for each approach and classifier. In addition, the second to last and last columns provide the weighted average for conversion Q_{wc} and the rank of an approach-classifier combination across all candidates in Table 3, respectively. Table 3 reveals that the overall best approach in the comparison is a response model with underlying ERT classifier ($Q_{wc} = 5{,}598$). This is a stunning result, suggesting that none of the uplift models outperforms a simple response model. Although the latter ignores the critical point that only persuadable customers are worth targeting, the incremental conversion of the response model exceeds that of the uplift approaches, which are deliberately designed to maximize incremental response. In this sense, the results of Table 3 put the merit of conversion uplift modeling very much into perspective.

The second-best approach in the comparison is LWUM developed on top of a random forest classifier ($Q_{wc} = 5{,}365$), followed by another implementation of this approach using the ERT classifier ($Q_{wc} = 5{,}316$). The other uplift approach, TCIA, performs much worse and proofs inferior to Lai's approach. LWUM was the overall best approach in a recent uplift modeling benchmark (Kane et al., 2014). In this sense, superiority over TCIA, which we observe, is consistent with prior work. However, the performance of the response modeling approach remains the key finding from the conversion modeling experiment. Delivering the largest incremental gain in conversions across all but the ninths decile, which is barely relevant for marketing practice, response modeling can well be considered a dominant approach for the employed data. This sets a hard benchmark for the revenue uplift experiment using the same data, which is presented in the next section.

Approach	**Model**	**Dec. 1**	**Dec. 2**	**Dec. 3**	**Dec. 4**	**Dec. 5**	**Dec. 6**	**Dec. 7**	**Dec. 8**	**Dec. 9**	Q_{wc}	**Rank**
Response	ERT	881	**1034**	**1231**	1261	1384	1656	1708	1673	1579	**5598**	**1**
	RFC	**916**	836	1197	**1333**	1363	1411	1451	1485	1505	5259	4
	LogR	913	973	759	1193	1171	1396	1304	1312	1459	4790	5
	KNN	299	550	940	1301	**1474**	1637	**1854**	**1860**	1742	4641	8
	SVM	190	225	413	585	1219	**1926**	1609	1649	1546	3339	11
	SGDC	168	427	636	777	921	1039	1216	1437	1551	3088	14
	NB	-206	214	442	623	1285	1197	1349	1358	1555	2622	17
LWUM	RFC	893	1034	1140	1258	1311	1439	1508	1639	1701	5365	2
	ERT	818	944	1070	1285	1428	1454	1582	1769	1807	5316	3
	LogR	855	951	909	998	1143	1180	1422	1436	1543	4676	7
	KNN	301	640	973	1241	1430	1574	1584	1622	1578	4510	9
	SGDC	185	422	606	763	981	1117	1263	1423	1559	3143	13
	SVM	-189	303	470	798	1189	1555	1621	1613	1584	3064	16
	NB	-243	286	408	232	871	1097	1282	1407	1533	2129	18
TCIA	GBC	775	843	905	1031	1281	1399	1438	1427	1579	4698	6
	SVM	-17	261	1033	1190	1281	1677	1578	1498	1685	3883	10
	KNN	288	490	694	854	916	880	1271	1457	1543	3286	12
	RFC	249	424	643	802	960	1025	1084	1157	1453	3087	15
	LogR	103	60	88	-96	399	962	1171	1573	**1922**	1587	19
	NB	11	229	84	64	463	710	965	1221	1838	1523	20

Table 3. Performance statistics of all approaches and models for conversion modeling.

7. Revenue Modeling

The proposed RDT is deployed as candidate for revenue modeling. To demonstrate its merits, it has been tested against TCIA and the benchmark of response revenue modeling.

Due to the nature of the RDT approach, i.e., the revenue response is a binary target variable after discretization, the models that have been presented for conversion modeling have been considered for predictions with this approach as well. Hence, next to 506 regression learners for response modeling and TCIA each, additional 316 classifiers have been considered on the RDT approach, making a sum of 1,328 models for the revenue experiment.

This section compares the performance of stated revenue models using the described performance measures. As before from the huge model library, only those models are considered that have passed parameter optimization with greatest success, i.e. each base learner's best model. Although the subsequent Qini curves visualize the per-decile performance of the underlying models as in conversion modeling, recall that the Qini value Q_r differs to Q_c in that it identifies the value instead of the number of incremental purchases. Figure 7 illustrates this value as a function of the respective population's fraction for the (i) revenue response transformation (top-left), (ii) response benchmark approach (top-right) and (iii) covariates transformation for revenue modeling (bottom-left). The legends display the model values of Q_r.

The performance of the shown models is summarized in Table 4 which reflects the results of the best models per decile and ranks them according to their weighted average for revenue. As before, for a theoretical fixed budget setting, the best approach and model combination is emphasized per decile.

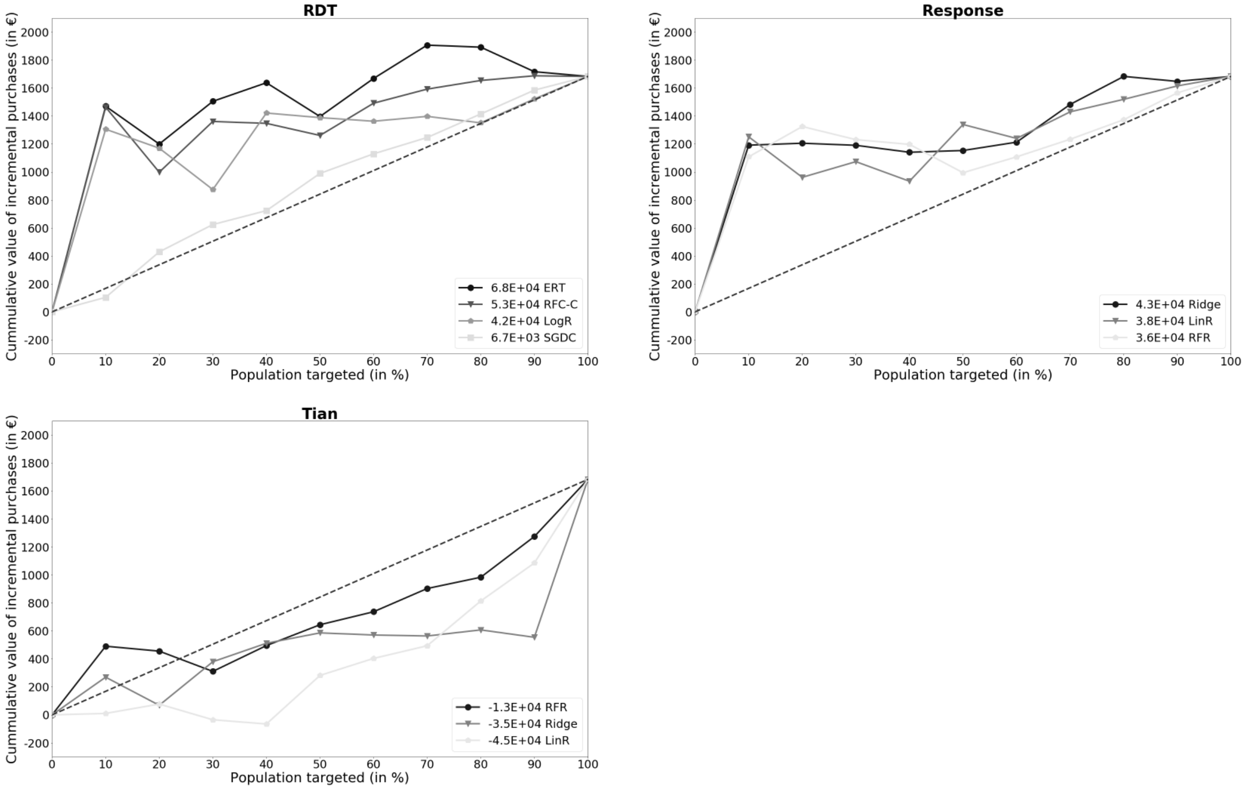

Figure 7. Best models per approach for revenue modeling.

Figure 7 and Table 4 clarify that the RDT approach outperforms response modeling and uplift covariates transformation on almost all deciles. At the whole, the best model on this approach, extremely randomized trees, ranks highest with $Q_{wr} = 6{,}806$€. Remarkably, when further comparing the performance of this model with all other models for each decile separately, it outperforms on eight out of nine deciles in total. Another interesting observation is that in terms of Q_{wr}, all models of RDT rank better than all response models which, in turn, dominate all transformation-related models (with one exception pointing out to SGDC). This further enhances the reliability of our claim that the proposed approach maximizes value not just occasionally in terms of a single model. TCIA not just performs worse compared to the other approaches; for the majority of deciles it not even complies with random targeting.

Approach	**Model**	**Dec. 1**	**Dec. 2**	**Dec. 3**	**Dec. 4**	**Dec. 5**	**Dec. 6**	**Dec. 7**	**Dec. 8**	**Dec. 9**	Q_{wr}	**Rank**
Response	Ridge	1191	1205	1191	1141	1153	1213	1483	1683	1647	5563	4
	RFR	1108	**1325**	1231	1197	995	1108	1234	1374	1565	5379	5
	LinR	1251	962	1074	934	1339	1239	1430	1519	1615	5267	6
RDT	ERT	**1470**	1199	**1505**	**1638**	**1396**	**1668**	**1907**	**1892**	**1717**	**6806**	**1**
	RFC-C	1463	998	1361	1347	1261	1491	1592	1654	1687	6081	2
	LogR	1305	1170	874	1421	1388	1363	1397	1351	1522	5656	3
	SGDC	104	429	624	724	990	1130	1246	1414	1585	3069	7
TCIA	RFR	491	456	311	495	644	737	903	984	1274	2533	8
	Ridge	269	69	380	512	586	571	564	607	555	1738	9
	LinR	14	77	-32	-65	282	403	496	813	1087	735	10

Table 4. Performance Statistics of all approaches and models for revenue modeling.

8. Comparison Conversion vs. Revenue Modeling

While the best models have been empirically examined for the conversion and revenue objective separately, the key question now refers to whether revenue uplift modeling provides more value compared to conversion uplift modeling and response modeling. This comparison is carried out in this section to not just demonstrate the superiority of revenue modeling for this type of marketing application and campaign, but to also prove the effectiveness of the proposed approach based on incremental revenue; a performance indicator being widely used in industry. Figure 8 and Table 5 present performances of the identified best model per conversion/revenue approach from the previous analyses.

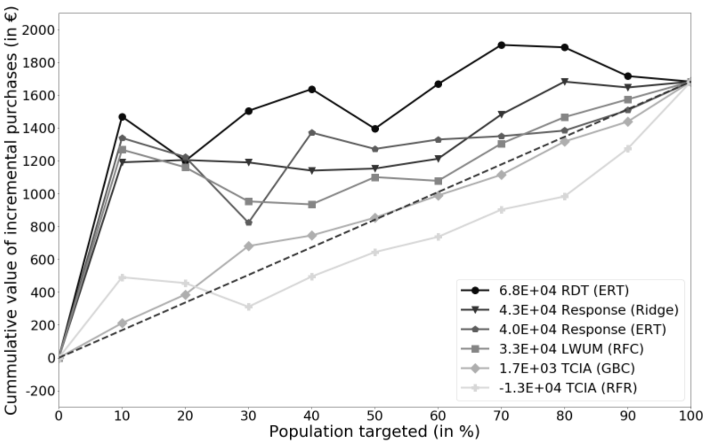

Figure 8. Top conversion and revenue models for incremental revenue.

From Figure 8 and Table 5 we learn that the extremely randomized trees learner on RDT is overall superior. Across all deciles, this model clearly outperforms (1) response modeling, (2) conversion uplift modeling (i.e., random forest classifier on LWUM and gradient boosting on TCIA) and (3) revenue uplift modeling in the shape of the random forest regressor that predicts with a transformed input space.

It is striking that of the models selected for this analysis, most of the top performers are tree-based and that, among them, ERT seems to be most valuable. Analyzing the best performance for each decile across objectives, the respective ERT models deliver the highest comparable value of incremental purchases on all deciles. The best model per decile is highlighted in bold font in Table 5.

Another argument in favor of the dominance of the RDT approach compared to the others stated stems from literature. Guelman (2014) suggests targeting the top ten percent most likely customers to respond positively to the campaign's treatment, i.e. only the customers from the first decile. Following this advice, RDT enhances incremental revenue comparably greatest with 1,470€. This is about 10% more incremental revenue than the second-best model as stated in Table 5.

The results confirm the contributions made in Bodapati and Gupta (2004) as discretizing revenue to apply classification models is a treasured possession we suggest campaign planners to carry in their toolboxes. According to the results of this paper, this is not just a valid but furthermore an innovative approach for extending the landscape of uplift modeling research and practice.

Approach	Objective	Model	Dec. 1	Dec. 2	Dec. 3	Dec. 4	Dec. 5	Dec. 6	Dec. 7	Dec. 8	Dec. 9
Response	Revenue	Ridge	1191	1205	1191	1141	1153	1213	1483	1683	1647
	Conversion	ERT	1340	**1224**	823	1372	1273	1330	1351	1385	1509
LWUM	Conversion	RFC	1268	1161	953	935	1101	1078	1305	1467	1575
RDT	Revenue	ERT	**1470**	1199	**1505**	**1638**	**1396**	**1668**	**1907**	**1892**	**1717**
TCIA	Revenue	RFR	491	456	311	495	644	737	903	984	1274
	Conversion	GBC	212	386	681	746	854	989	1115	1318	1439

Table 5. Incremental revenue of best conversion and revenue models.

9. Conclusion

Empirical results have confirmed the proposed approach to be a valuable tool for revenue uplift modeling. For the data at hand, the parameter-optimized extremely randomized tree algorithm on RDT is most successful in identifying persuadable customers based. In other words, compared to other approaches considered in the comparison, RDT achieves the largest increase in incremental revenue. Although the model's uplift estimate is not unbiased, model building on a discretized (i.e., binary) response implies a smaller bias compared to an unmodified, continuous revenue response (Bodapati and Gupta, 2004).

More generally, the paper has reviewed several uplift modeling approaches and compared their effectiveness against each other and traditional response modeling in a large-scale experiment. Experimental results suggest that uplift modeling does not outperform response modeling in terms of conversion, whereas revenue uplift modeling does add value. Accordingly, the proposed approach complements previous uplift modeling strategies and provides better performance when targeting marketing campaigns the primary goal of which is increasing revenue. Next to the comprehensive empirical study, the paper has developed a formalized uplift modeling process for marketing.

The applicability of the proposed model is not restricted to the online sphere. In fact, the original idea of uplift modeling stems from an offline setting. Many authors have indicated the effectiveness of uplift modeling with physical marketing incentives. These include Guelman et al. (2012, 2015) who sent out information letters and conducted outbound courtesy calls within the insurance industry, Kane et al. (2014) who point to a direct paper mail campaign and Radcliffe (2007) who uses catalogue mails in retail. If the data requirements for uplift modeling are fulfilled (i.e., random assignment of customers to the treatment group and sufficient number of samples), offline retailers such as brick-and-mortar stores can also make use of the approaches and models described here. There also exist situations where online communications take place (e.g., per e-mail), but purchases are undertaken offline (e.g. in brick-and-mortar stores), which build a bridge between online and offline interactions.

As indicated with the UMPM, revenue uplift models should be considered only if the marketing goal is to maximize incremental revenue. By comparing revenue uplift models to conversion uplift models and response models, we empirically confirmed that the former is superior if the campaign goal is revenue maximization. We may thus recommend targeting corresponding campaigns using the modeling approach proposed here. However, if customer acquisition/retention is the primary marketing goal, previous uplift approaches are probably better suited.

In future research, the discretization of the revenue response could be modified in that not a binary variable is induced but a categorical one (i.e., coarsening revenue) converting it into a multi-class classification problem. This would be especially valuable to account for broad or multimodal distributions of customer spending. Furthermore, in terms of the generalizability of RDT, it would be interesting to examine application areas other than e-couponing. A final note for future research is directed to design direct revenue uplift models that are out of the scope of transformation-based modeling architectures.

References

Baumann, A., Lessmann, S., Coussement, K. and De Bock, K. W. 2015. "Maximize What Matters: Predicting Customer Churn with Decision-Centric Ensemble Selection," in *Proceedings of the 23rd European Conference on Information Systems* (ECIS'15), Münster, Germany.

Bodapati. A., and Gupta. S. 2004. "A Direct Approach to Predicting Discretized Response in Target Marketing," *Journal of Marketing Research* (41:1), pp. 73-85.

Bose, I., and Xi, C. 2009. "Quantitative Models for Direct Marketing: A Review from Systems Perspective," *European Journal of Operational Research* (195:1), pp. 1-16.

Chen, Z.-Y., Fan, Z.-P., and Sun, M. 2015. "Behavior-Aware User Response Modeling in Social Media: Learning from Diverse Heterogeneous Data," *European Journal of Operational Research* (241:2), pp. 422-434.

Chickering, D. M., and Heckerman, D. 2000. "A Decision Theoretic Approach to Targeted Advertising," in *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, pp. 82-88.

eMarketer. 2016. "Worldwide Ad Spending: eMarketer's Updated Estimates and Forecast for 2015-2020," (available at http://www.strathcom.com/wp-content/uploads/2016/11/eMarketer_Worldwide_Ad _Spending-eMarketers_Updated_Estimates_and_Forecast_for_20152020.pdf; retrieved May 5, 2017).

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. 1996. "The KDD Process for Extracting Useful Knowledge from Volumes of Data," *Communications of the ACM* (39:11), pp. 27-34.

Fernandes, R., and Leblanc, S. G. 2005. "Parametric (Modified Least Squares) and Non-Parametric (Theil–Sen) Linear Regressions for Predicting Biophysical Parameters in the Presence of Measurement Errors," *Remote Sensing of Environment* (95:3), pp. 303-316.

Gotz, M., Weber, C., Blocher, J., Stieltjes, B., Meinzer, H. P., and Maier-Hein. K. 2014. "Extremely Randomized Trees Based Brain Tumor Segmentation." in *Proceedings of BRATS Challenge-MICCAI*.

Guelman, L., Guillén, M., and Pérez-Marín, A. M. 2012. "Random forests for uplift modeling: an insurance customer retention case," in *Modeling and Simulation in Engineering, Economics and Management*, Springer Berlin Heidelberg, pp. 123-133.

Guelman, L. 2014. "Optimal Personalized Treatment Learning Models with Insurance Applications," Doctoral Thesis, Universitat de Barcelona, Barcelona, Spain.

Guelman, L., Guillén, M., and Pérez-Marín. A. M. 2015. "Uplift Random Forests," *Cybernetics and Systems* (46:3-4), pp. 230-248.

Hansen, B. B., and Bowers, J. 2008. "Covariate Balance in Simple. Stratified and Clustered Comparative Studies," *Statistical Science* (23:2), pp. 219-236.

Hansotia, B., and Rukstales, B. 2002. "Incremental Value Modeling," *Journal of Interactive Marketing* (16:3), pp. 35-46.

Hastie, T., Tibshirani, R., and Friedman. J. 2009. *The Elements of Statistical Learning*, Second Edition. New York: Springer.

Holland, P. W. 1986. "Statistics and Causal Inference," *Journal of the American Statistical Association* (81:396), pp. 945-960.

Hua, S. 2016. "What Makes Underwriting and Non-Underwriting Clients of Brokerage Firms Receive Different Recommendations? An Application of Uplift Random Forest Model," *International Journal of Finance and Banking Studies (2147-4486)* (5:3), pp. 42-56.

Imbens, G. W. 2004. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *The Review of Economics and Statistics* (86), pp. 4-29.

Jacobs, B. J. D., Donkers, B., and Fok, D. 2016. "Model-Based Purchase Predictions for Large Assortments," *Marketing Science* (35:3), pp. 389-404.

Jaroszewicz, S. 2016. *First Ideas on Revenue Uplift Modeling*, Personal Communication, October 15, 2016.

Jaroszewicz, S., and Zaniewicz, Ł. 2016. "Székely Regularization for Uplift Modeling," in *Challenges in Computational Statistics and Data Mining*, Springer International Publishing, pp. 135-154.

Jaskowski, M., and Jaroszewicz, S. 2012. "Uplift Modeling for Clinical Trial Data," *ICML 2012 Workshop on Clinical Data Analysis*.

Kane, K., Lo, V. S., and Zheng. J. 2014. "Mining for the Truly Responsive Customers and Prospects Using True-Lift Modeling: Comparison of New and Existing Methods," *Journal of Marketing Analytics* (82:4), pp. 218-238.

Khajehzadeh, S., Oppewal, H., and Tojib, D. 2014. "Consumer Responses to Mobile Coupons: The Roles of Shopping Motivation and Regulatory Fit," *Journal of Business Research* (67), pp. 2447-2455.

Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., and Pohlmann, N. (2013). "Online controlled experiments at large scale," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1168-1176), I. S. Dhillon, Y. Koren, R. Ghani, T. E. Senator, P. Bradley, R. Parekh, J. He, R. L. Grossman & R. Uthurusamy (eds.), Chicago, IL, USA: ACM.

Kohavi, R., Longbotham, R., Sommerfield, D., and Henne, R. M. 2009. "Controlled Experiments on the Web: Survey and Practical Guide," *Data Mining and Knowledge Discovery* (18:1), pp. 140-181.

Lai, L. T. 2006. "Influential Marketing: A New Direct Marketing Strategy Addressing the Existence of Voluntary Buyers," Master of Science Thesis, Simon Fraser University School of Computing Science. Burnaby, BC, Canada.

Larsen, K. 2010. *Net Lift Models*. Slides of a Talk at the A2010 Analytics Conference. September 2-3 2010. Copenhagen. Denmark.

Ling, C. X., and Li, C. 1998. "Data Mining for Direct Marketing: Problems and Solutions," *KDD (Vol. 98)*, pp. 73-79.

Lo, V. S. 2002. "The True Lift Model: A Novel Data Mining Approach to Response Modeling in Database Marketing," *ACM SIGKDD Explorations Newsletter* (4:2), pp. 78-86.

Lo, V. S., and Pachamanova, D. A. 2015. "From Predictive Uplift Modeling to Prescriptive Uplift Analytics: A Practical Approach to Treatment Optimization While Accounting for Estimation Risk," *Journal of Marketing Analytics* (2), pp. 79-95.

Long, J. S. 1994. "Confirmatory Factor Analysis: A Preface to LISREL," in *Factor Analysis and Related Techniques,* M. S. Lewis-Beck (eds.), London: Sage Publication.

Manahan, C. 2005. "A Proportional Hazards Approach to Campaign List Selection," *SAS User Group International (SUGI) 30 Proceedings*.

Media Dynamics, Inc. 2014. *America's Media Usage & Ad Exposure: 1945-2014*, September 21 2014, Nutley, New Jersey.

Nassif, H., Kuusisto, F., Burnside, E. S., and Shavlik, J. W. 2013. "Uplift Modeling with ROC: An SRL Case Study," in *Proceedings of the International Conference on Inductive Logic Programming (ILP'13)*, Rio de Janeiro, Brazil, pp. 40-45.

Nattee, C., Khamsemanan, N., Lawtrakul, L., Toochinda, P., and Hannongbua, S. 2017. "A Novel Prediction Approach for Antimalarial Activities of Trimethoprim, Pyrimethamine, and Cycloguanil Analogues Using Extremely Randomized Trees," *Journal of Molecular Graphics and Modelling* (71), pp. 13-27.

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., and Mason, C. H. 2006. "Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models," *Journal of Marketing Research* (43:2), pp. 204-211.

Netessine, S., Savin, S., and Xiao, W. 2006. "Revenue Management Through Dynamic Cross Selling in E-Commerce Retailing," *Operations Research* (54:5), pp. 893-913.

Ngai, E. W. T., Xiu, L., and Chau, D. C. K. 2009. "Application of Data Mining Techniques in Customer Relationship Management: A Literature Review and Classification," *Expert Systems with Applications* 36(2), pp. 2592-2602.

Park, C. H., and Park, Y.-H. 2016. "Investigating Purchase Conversion by Uncovering Online Visit Patterns," *Marketing Science* (35:6), pp. 894-914.

Platt, J. 1999. "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods," *Advances in Large Margin Classifiers* (10:3), pp. 61-74.

Radcliffe, N. J. 2007. "Using Control Groups to Target on Predicted Lift: Building and Assessing Uplift Models," *Direct Marketing Journal, Direct Marketing Association Analytics Council* (1), pp. 14-21.

Radcliffe, N. J., and Surry, P. D. 1999. "Differential Response Analysis: Modeling True Response by Isolating the Effect of a Single Action," *Credit Scoring and Credit Control VI*.

Radcliffe, N. J., and Surry, P. D. 2011. *Real-World Uplift Modelling with Significance-Based Uplift Trees*, White Paper TR-2011-1. Stochastic Solutions.

Reinartz, W. J., and Kumar, V. 2003. "The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration," *Journal of Marketing* (67:1), pp. 77-99.

Rzepakowski, P., and Jaroszewicz, S. 2012a. "Uplift Modeling in Direct Marketing," *Journal of Telecommunications and Information Technology*, pp. 43-50.

Rzepakowski, P., and Jaroszewicz, S. 2012b. "Decision Trees for Uplift Modeling with Single and Multiple Treatments," *Knowledge and Information Systems* (32:2), pp. 303-327.

Siegel, E. 2011. *Uplift Modeling: Predictive Analytics Can't Optimize Marketing Decisions Without It*, Prediction Impact White Paper sponsored by Pitney Bowes Business Insight.

Siegel, E. 2013. *Predictive analytics: The power to predict who will click, buy, lie, or die*, John Wiley & Sons.

Soltaninejad, M., Yang, G., Lambrou, T., Allinson, N., Jones, T. L., Barrick, T. R., and Ye, X. 2016. "Automated Brain Tumour Detection and Segmentation Using Superpixel-Based Extremely Randomized Trees in FLAIR MRI," *International Journal of Computer Assisted Radiology and Surgery*, pp. 1-21.

Sołtys, M., Jaroszewicz, S., and Rzepakowski, P. 2015. "Ensemble Methods for Uplift Modeling," *Data Mining and Knowledge Discovery* (29:6), pp. 1531-1559.

Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. 2014. "A Simple Method for Estimating Interactions Between a Treatment and a Large Number of Covariates," *Journal of the American Statistical Association* (109:508), pp. 1517-1532.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., and Baesens, B. 2012. "New Insights into Churn Prediction in the Telecommunication Sector: A Profit Driven Data Mining Approach," *European Journal of Operational Research* (218:1), pp. 211-229.

Yong, F. 2015. "Quantitative Methods for Stratified Medicine," Doctoral Thesis, Harvard University, Graduate School of Arts & Sciences, Cambridge, Massachusetts.

Zaniewicz, Ł., and Jaroszewicz, S. 2013. "Support Vector Machines for Uplift Modeling," *Data Mining Workshops (ICDMW). 2013 IEEE 13th International Conference*, pp. 131-138.

3 Organizational Perspective: User Behavior in Network Structures

3.1 User Behavior in Social Media

ARTICLE 11:

THE ROLE OF GENDER IN BLOGGING: CURRENT STATE OF RESEARCH AND FUTURE DIRECTIONS

Authors: Annika Baumann, Humboldt-Universität zu Berlin, Berlin, Germany

Hanna Krasnova, Universität Potsdam, Potsdam, Germany

Natasha Veltri, University of Tampa, Tampa, Florida, USA

Publication status: Working Paper

Abstract

Blogging is one of the most popular forms of online communication, enabling individuals to share their personal ideas, knowledge, and thoughts with the global community. The blogosphere has evolved from the domain for mainly personal use to a powerful tool to disseminate product reviews, political views, and even form public opinion. While blogging is equally popular among men and women, top influential bloggers are most exclusively male. To contribute to the gender parity in the blogosphere we conduct a comprehensive meta-review of past studies on blogging to understand gender differences and similarities in this context. Structured with regard to the type of blogging, bloggers' motivations, writing style and characteristics, their attitudes to privacy, blogging content, and blog readership, our findings provide evidence for the emergence of new ways in which this technology helps users to transgress gender boundaries. Based on our findings, we define perspective research areas for future investigations.

Keywords: *Blog(s), Blogging Community, Literature Review, Gender Differences, User Behavior*

1. Introduction

Over the last twenty years advances in information and communication technology have enabled dynamic and rich online communication supporting new types of social structures (Butler, 2001). Anyone with access to the Internet became a member of the online community, and both a consumer and a contributor to the constantly evolving social web. In this global online environment individuals can share their personal ideas, opinions, thoughts, and stories through personal journaling or blogging, one of the earliest and most popular forms of online communication (Chau and Xu, 2012). Blogging has experienced tremendous growth worldwide with the number of blogs increasing from 35.8 million in 2006 to 181 million in 2011 (Nielsen, 2012). Initially blogs were mainly used to express personal opinions, document experiences, and showcase interests. However, as the blogosphere evolved, blogging became not only a tool for individual musings, but also a potent avenue for a variety of professional and semi-professional uses. Increasingly, blogs are used to disseminate product reviews, promote products, services or political views, and even form public opinion. In fact, consumers rank blogs among the top five most trustworthy sources of information on the Internet and third most influential digital resource when it comes to purchasing products and services (TechnoratiMedia, 2013). In the light of these developments, it comes as no surprise that the most successful bloggers have become much sought out influencers.

Blogging is equally popular among men and women (Sysomos, 2010) and an average top blog has 45% female and 55% male readers (Pingdom, 2013). Thus, on the one hand, there is evidence of gender parity in the blogosphere. On the other hand, however, the most influential bloggers continue to be the Internet veterans who are usually white males, indicating that online blogging community follows existing gender patterns (Pedersen and Macafee, 2007). This gender bias among influencers is troublesome because it further contributes to existing stereotypes, skews the public sentiment by inundating female voices, and even puts female bloggers at an economic disadvantage. Indeed, successful bloggers earn income for themselves from banner and text advertising, affiliate programs, sponsored content and product reviews. Businesses and society can also benefit from a more prominent role of women in the blogosphere by both featuring female opinions on their products and services and attracting female readership.

Several possible factors could be responsible for existing differences. While there is a debate as to the existence of the gender digital divide when it comes to the use of technology (e.g. Joiner et al., 2005; Hilbert, 2011), there are well-documented differences in how women and men adopt and use technology and the Internet (Venkatesh and Morris, 2000; Jackson et al., 2001). Many of these differences are rooted in relationship orientation typical for women and a more goal-oriented approach characteristic of men (Gefen and Straub, 1997; Teo and Lim, 2000). Further, social norms often dictate gender roles. Because much of the gender association is socially constructed (Eagly, 1987; Eagly and Steffen, 1984), existing stereotypes play into how society perceives editorials and opinion blogs written by men and women and also impact the topics that female and male bloggers choose. Furthermore, poor quality of female blogs, unwillingness of top bloggers to link to them and lack of interest on the part of female bloggers to post about broader topics have all contributed to gender disparity particularly in the political blogosphere (Harp and Tremayne, 2006). All in all, however, understanding of these developments is hindered by the lack of a comprehensive perspective on gender differences between bloggers. While multiple studies report various gender-relevant findings, these insights are scattered and unstructured.

As the blogosphere continues to evolve and become more powerful, understanding the differences in men and women's use of this technology is essential for ensuring gender parity and a globally equitable information society (Hafkin and Huyer, 2007). We contribute to this goal by conducting a comprehensive meta-review of past studies to paint a more complete picture and identify critical knowledge gaps. We structure existing insights on gender differences with regard to the type of blogging, bloggers' motivations and characteristics, their attitudes to privacy, content and style of blogs, and blog readership. Our findings provide a deeper insight into the underlying dynamics of gender differences, help to identify ways to improve the disparity and give impetus to future research efforts in this area.

The remainder of the paper is structured as follows. First, we present some theoretical background regarding blogging and gender differences in general. Chapter 3 will describe the methodological approach we applied for our comprehensive literature survey. Chapter 4 will present the gender findings in the blogosphere, systematically classified into their relevant categories. In the following, results are coherently discussed and directions for future research presented. Finally, we discuss the limitations of our research.

2. Theoretical Background

The chapter provides the theoretical background for the main concepts of the paper. Chapter 2.1 provides insights on blogging in general, providing definitions and typical characteristic of blogs. The following chapter presents theories with respect to gender differences in the online and offline setting.

2.1. Blogging

With top blogging platforms, such as Blogger, WordPress.com, and Tumblr, attracting a whopping 80 million unique visitors a month (Nielsen, 2012), the influence of blogs on individuals, businesses and society is unprecedented. Herring et al. (2004a, p.1) provide a broad definition of blogs, describing them as "frequently modified web pages in which dated entries are listed in reverse chronological sequence". Kelleher and Miller (2006) complement this view by underlining such important features of blogs as readers' ability to comment and the presence of hyperlinks (e.g. to other blogs) (Cavanaugh, 2002). Various types of blogs exist, differing in terms of their primary purpose and other characteristics. First, blogs can be used both personally and professionally. While this distinction is sometimes blurred, personal use implies blogging from an individual and typically independent perspective, and professional use involves endorsements or opinions on the part of a blogger clearly affiliated with a specific organization or brand (Sifry, 2004; Kelleher and Miller, 2006).

2.2. Gender Differences in Offline and Online Settings

So far, multiple theories have tried to address the origin and implications of gender differences, offering various perspectives on this phenomenon. Among them, Bem's (1981) gender schema theory suggests that gender beliefs are formed in early childhood by cultural expectations. The evolutionary psychology uses a similar theoretical approach and explains gender differences by human ancestral past which over thousands of years fostered women to be more compassionate and men to more competitive (Stewart-Williams and Thomas, 2013). Other theories have also gained in importance, including Cross and Madson's self-construal theory (1997). Based on this theory, men favor independence, whereas women are more interdependent. Hence, men strive to remain autonomous when it comes to their relationships with others, while women rather prefer closeness and intimacy in their relationships. Building on and extending the views of Cross and Madson (1997), Baumeister and Sommer (1997) argued that both men and women have the fundamental need to belong in terms of relationships to others, but they accomplish it by using different strategies. In general, women focus on rather close, two-way relationships with high intimacy, whereas men navigate in a high status hierarchy within broader groups (Baumeister &

Sommer, 1997). Together, these differences between men and women are likely to underlie the distinct patterns in their attitudes, behaviors and roles they play in the society.

3. Methodology

For our meta-review we followed the recommendations of Webster and Watson (2002) and Levy and Ellis (2006). Initially, a keyword search was performed in the databases JSTOR, ScienceDirect, EBSCOhost, Springer, Wiley Online Library, Emerald Insight and GoogleScholar. We used combinations of the following keywords across all databases: {Blog, Blogging} and {gender, men, women, female, male, boy, girl}. We did not limit the search by the year of publication, and included only peer-reviewed studies written in English. No other search constraints or filters were used. Since our focus was on documenting gender differences in blogging, we reviewed the title and abstract of all search result to identify and select potentially relevant articles. Those pre-selected articles were further scanned using the in-text search for relevant markers such as “women”, “woman”, “female”, “male”, “men”, “man”, “boy”, “girl” to find potentially relevant findings with respect to gender differences on different dimensions of blog usage behavior. For our analysis we only selected work that explicitly addressed differences between men and women with regard to blogging based on original empirical results, whereas articles that dealt with only female or male perspectives were used as supporting evidence (e.g. Simmons, 2008; Somolu, 2007). Articles which contained such findings were finally classified as being relevant for our meta-review. The second step involved a backward search, where we reviewed relevant citations from those already identified articles. Finally, an additional forward search based upon the already identified articles was conducted using Google Scholar.

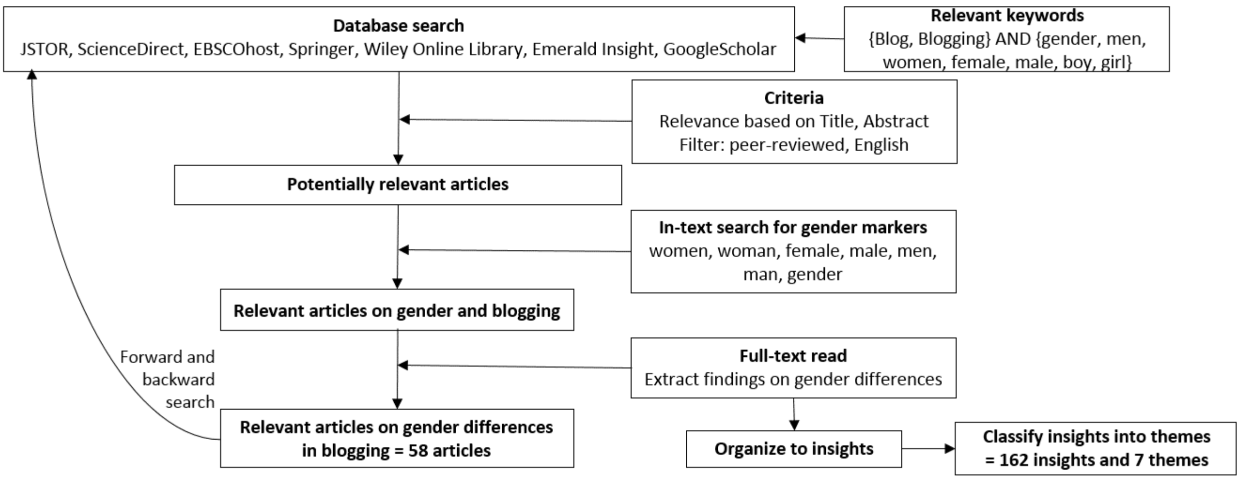

Figure 1. Process of identification of relevant articles and extraction of insights.

The final sample included 58 studies published between 2004 and 2017. 77.6% of articles appeared in journals, and 15.5% in conferences[15]. 30 different journals published gender-relevant insights on blogging, underscoring tremendous interest of this subject to a variety of scientific communities, including information systems, journalism, and psychology, among others. Nine articles appeared in the “Journal of Computer-Mediated Communication”, four in “Computers in Human Behavior”, three in the “Proceedings of the AAAI Conference”. Furthermore two articles appeared in the journals “Cyberpsychology and Behavior” and “New Media & Society” respectively. 55.2% of articles specifically focused on gender aspects of blogging. Student samples were not common - only 9 studies used students as a source of data. Most often real data extracted from blogs was used (62.1%), with

[15] Complete list of examined articles and journals is available from authors upon request.

content analysis being the main method for analysis. Surveys were the second most important method of data collection (34.5% of included articles) with a broad range of techniques used for their evaluation (e.g. PLS, multivariate analysis, and descriptive statistics). Finally, because we were only able to review articles in English, most of the studies we identified relied on data collected in North America using English language blogs.

All articles were thoroughly reviewed and relevant findings were documented. Since studies often contained heterogeneous findings, we organized them into smaller chunks of information with singular meaning – *insights*. For example, the finding "the use of 'navy', 'gold', and 'silver' were most telling of masculine writing; the use of 'purple', 'tan', and 'pink' were telltale of feminine writing" (Liu and Mihalcea, 2007) is parted into the insights "females use the words purple, tan, pink" and "males use the words navy, gold, silver". In total 162 insights were identified. In the next step two authors have independently reviewed those insights in an effort to identify a set of leading themes. Their initial suggestions were then compared and discussed by all four authors of the study. This approach uncovered seven themes dominant in the gender discourse on blogging (Table 1).

Theme	Theme Description	Number and relative share of insights
Motivation to Blog	Insights on the role of gender in motivational patterns of bloggers.	22 (13.6%)
Types of Blogs	Insights on the role of gender in the choice of a blog type (e.g. filter, diary blogs, etc.).	14 (8.6%)
Bloggers' Characteristics	Characteristics of the blog owner viewed from the gender perspective.	21 (13.0%)
Blog Content	Insights on the role of gender in the choice of the blog theme and specific content.	34 (21.0%)
Readership	Insights on the role of gender in the attention attracted by the blog and readers' perceptions of it.	9 (5.6%)
Writing Style	Insights on the writing style and word choice of blog entries and comments from the gender perspective.	39 (24.1%)
Privacy	Insights on the role of gender in privacy perceptions, privacy behaviour and past privacy experiences of bloggers.	23 (14.2%)

Table 1. Themes in gender-relevant discourse on blogging.

In the next step, each insight was assigned to a specific theme by two coders. Inter-coder reliability measured by Cohen's kappa was high, reaching 0.747, providing evidence of a high level of agreement between both coders (Landis and Koch, 1977). The final decision on the distribution of themes was reached by consensus.

4. Results: Gender Differences in Blogging

This chapter presents insights collected through literature survey with respect to the themes defined in Chapter 3.

4.1. Motivation to Blog

Initial motives to blog are likely to play a defining role in how the blogging functionality is used (Kim et al. 2007; Deci, 1975). Our review of past non-gender-specific research reveals that a variety of factors, such as social connection, self-expression, documenting, enjoyment, as well as attention and reputation may motivate bloggers (Table 2). However, the motivational strength of many determinants is ambiguous, with research reporting conflicting results. Notable exceptions are motives of self-expression, documenting and enjoyment, for which only significant associations with markers of blogging participation have been reported. Ambiguity in reported findings can be traced back to the differences in the blogging types and platforms explored in particular studies and to the heterogeneity of the examined samples, especially with respect to the gender variable.

Motive/Study	**Social Connection**	**Self-Expression**	**Attention and Reputation**	**Documenting**	**Information Hub**	**Pass Time**	**Feedback**	**Professional**	**Altruism**	**Enjoyment**	**Perceived Usefulness**	**Information Seeking**	**Emotional Outlet**	**Advertising**
Hollenbaugh (2010)	○		○	●	○	○	○	○	○					
Hsu and Lin (2008)	○		●				○		●	●	○			
Lu and Hsiao (2009)		●	●											
Liao et al. (2011)	●	●	●	●		●	●			●				
Chen (2012)	●	●	○											
Huang et al. (2007)	●	●		●	●							●		
Hollenbaugh (2011)	✓		✓	✓	✓	✓	✓	✓	✓					
Trammell et al. (2006)	✓	✓			✓	✓		✓		✓				
Li (2005)	✓	✓		✓	✓	✓					✓			
Nardi et al. (2004)	✓	✓		✓	✓									
Fullwood et al. (2015)		✓											✓	✓

Table 2. *Motives for blogging explored in past studies (● (○) – this motive was (not) significantly associated with a tested marker of blogging participation; ✓- this motive was established in the course of a qualitative study).*

Only a limited number of studies report gender differences in the motivational structure of bloggers (Table 3). With some degree of generalization, we observe that most blogging motives are consistent with known gender-specific motivational patterns in communication and technology use (Furman and Buhrmester, 1985; Tannen, 1994). Social connection plays a greater role for female bloggers (e.g. Hollenbaugh, 2010), and self-expression and documenting serve as additional motives to blog (e.g. Li, 2005), with some studies, however, finding no difference compared to males (e.g. Hollenbaugh, 2011). In contrast, male bloggers emphasize professional uses of blogging, such as finding a better job (Hollenbaugh, 2010). Attention and reputation motive as well as the use of blogging to broadcast one's ideas also motivate male bloggers, even though evidence remains limited and ambiguous.

Motive/Study	**Social Connection**	**Self-Expression**	**Attention and Reputation**	**Documenting**	**Information Hub**	**Professional**	**Pass Time**	**Feedback**	**Enjoyment**	**Perceived Usefulness**
Hollenbaugh (2010)	**W**	○		**W**	○	**M**	○	○		
Hollenbaugh (2011)	**W**	○		○	○	**M**	○	○		
Trammell et al. (2006)*	**W**	○			**M**	**M**	○		○	
Lu and Hsiao (2009)		**W**	**M**							
Li (2005)	○	**W**		**W**	**M**		**W**			○

Table 3. *Gender differences in motivational patterns of bloggers (*statistical significance was not reported. **W** – this motive is stronger for women; **M** –stronger for men; ○ – no difference).*

All in all, female bloggers are motivated by the opportunities to support social relationships, outpour and document their personal narratives. Men, on the other hand, rather adopt a more goal-oriented approach when blogging, striving to attract attention and develop themselves professionally. This, however, does not imply that women do not pursue these motives at all. For example, personal journaling has been linked to greater autonomy and control for women (Stavrositu and Sundar, 2012), indicating that blogging could serve as a tool to empower women. Furthermore, certain personality factors might influence the motivation of individuals to blog such as males who score rather low on openness are more likely to use blogs as an emotional outlet (Fullwood et al., 2015).

Even though past findings identify certain patterns, available evidence is surprisingly scarce and in many cases inconclusive, calling for more research in this area. For example, motives such as getting feedback, altruism, enjoyment, perceived usefulness and information seeking activities – identified as important in the past non-gender specific research (see Table 2) - have received no or little attention.

4.2. Types of Blogs

In line with the differences in motivational patterns established above, our review reveals a relatively consistent picture with regard to male and female preferences for certain blog types. It appears that men have a tendency to operate filter, knowledge and mixed blogs (Fullwood et al., 2013; Herring et al., 2004a; Herring et al., 2004b; Herring et al., 2005a; Kennedy et al., 2005; Trammell and Keshelashvili, 2005), while women have a preference for personal / diary / journal blogs (Fullwood et al., 2013; Herring et al., 2004a; Herring et al., 2004b; Herring et al., 2005a; Kennedy et al., 2005; Wei, 2009; Nowson and Oberlander, 2006; Trammell and Keshelashvili, 2005; Viégas, 2005). These findings are consistent with the traditional view on men and women in the offline context (Tannen, 1994) as well as the recent self-construal theory (Cross and Madson, 1997): While filters and knowledge blogs are primarily used to disseminate factual knowledge, personal blogs often focus on personal experiences of the blogger and are therefore, more emotional. However, there are some recent developments, which imply that existing, traditional gender lines may begin to blur in the blogosphere. For example, Van Doorn et al. (2007) find that the feminine, diary-oriented blogging is adopted by both sexes, which suggests that blogs provide room for the “female” voice on the Internet. Furthermore, women who set high standards for themselves in terms of achievement exhibit a tendency to operate filter blogs (Chen, 2012), and hence strive to assert themselves as information hubs alongside male bloggers. All in all, it appears that blogs may have an empowering effect on women (Stavrositu and Sundar, 2012), even though these trends still need time to mature.

4.3. Bloggers’ Characteristics

To learn about blogosphere’s gender composition and to describe different qualities of bloggers (e.g., in terms of age, effort, tendencies to cooperate) as a function of gender, we collected existing insights on their gender characteristics (Table 4).

One the one hand, men may be more likely to maintain a blog because they are more technically sophisticated in the context of web (Christie, 2005) and the Internet is traditionally seen as a male domain (Harp and Tremayne, 2006). Indeed, male bloggers have more blogging experience (Miller et al., 2011) and older bloggers are predominantly male (Schler et al., 2006). On the other hand, more women than men seem to be active in the blogosphere (e.g. Carstensen, 2009; Schler et al., 2006; Blinka and Smahel, 2009). Indeed, women are more likely to operate personal journal type blogs, which are the most popular blog type on the Internet (Herring et al., 2004a). Further, the emergence of blogging services, like Wordpress and Tumblr, has made it possible to create a blog with a few simple steps and with only limited technical knowledge. This made blogging feasible even for technically unsophisticated individuals, especially women who tend to stick to basic default settings for their blogs (Mazur and Kozarian, 2010).

When it comes to personal characteristics of bloggers, we find some evidence that conflicts male and female behavior offline. In contrast to traditional gender roles emphasizing women’s relationship-orientation, in the blogosphere men cooperate with other bloggers more frequently (Miller et al., 2011), mention others more often (Herring et al., 2005b) and stay in contact with others more using RSS feeds (Lu et al., 2010). However, women are more concerned about who visited their blog and use friends lists and visiting card functions (Lu et al., 2010), suggesting an interest in their readership and promotion of their blog (which is more in line with professional and goal-oriented male bloggers). Moreover, we observe that women also spend more time (Child et al., 2012) and effort maintaining their blogs (Nowson and Oberlander, 2006). On the one hand, this finding confirms that women are more concerned about the opinions of others (Eagly, 1978). Alternatively, however, this may indicate that female bloggers are as much interested in the success of their blogs as their male counterparts, suggesting a rather competitive nature of female blog writers – a characterization typically reserved for males (Cross and Madson, 1997).

Domain	Gender	Insights	Source
Blogosphere Composition	Women	Young bloggers are predominantly female	Schler et al. (2006); Herring et al. (2004a)
		More female authors of blogs	Carstensen (2009) Liu (2017)
		Blogs are more attractive for girls	Blinka and Smahel (2009)
	Men	Older bloggers are predominantly male	Schler et al. (2006); Herring et al. (2004a)
		More blogging experience (3.6 vs. 3.2 years)	Miller et al. (2011)
	Both	Both use blogs equally as often	Porter et al. (2009)
Personal Traits	Women	Women with high neuroticism are more likely to blog	Guadagno et al. (2008)
		Want to know who visited their blogs, use friend lists, visiting card functions	Lu et al. (2010)
	Men	No difference of likelihood to blog based on level of neuroticism	Guadagno et al. (2008)
		Stronger flow experience while blogging	Lu et al. (2010)
Effort	Women	Women spent more time maintaining their blog	Child et al. (2012)
		Young female bloggers have more technical advanced blogs.	Mazur and Li (2016)
		Invest more efforts in blogging (frequency, length)	Nowson and Oberlander (2006)
	Both	No difference in posting frequency	Miller et al. (2011)
Cooperation	Men	Use RSS to keep up with other bloggers	Lu et al. (2010)
		More often blog with others	Miller et al. (2011)
		More mentions of other bloggers	Herring et al. (2005b)

Table 4. Gender-relevant findings on blogger characteristics.

4.4. Blog Content

In their offline communication women are known to be more supportive, social, polite, and expressive (Cross and Madson, 1997); emphasize feelings (Haas, 1979) and exhibit empathy (Clark and Bittle, 1992). Topic-wise, they are more likely to discuss home and family (Haas, 1979), and focus on subjective aspects (Broverman et al., 1968). In contrast, males are more attentive to the overall message theme (Meyers-Levy, 1989), and rather communicate facts and information (Tannen, 1994). Finding themselves under higher pressure to present their accomplishments and establish social standing (Tannen, 1994), men are more likely to self-assert themselves by communicating symbols of their success (Gefen and Ridings, 2005), and hiding their real self (Jourard, 1971, p. 35). These stereotypes are deeply engrained in the social consciousness, positioning women as family-focused and talkative; while men are expected to be assertive but reserved (Eagly and Steffen, 1984).

Focus		Findings	Source
Relationships to Others and Society	Women	Family, non-romantic peer relationships, loved ones, romance	Mazur and Kozarian (2010); Liu and Mihalcea (2007); Argamon et al. (2007); Garimella and Mihalcea (2016)
		Personal content, lives, thoughts, concerns, and feelings	Pedersen and Macafee (2007); Schler et al. (2006); Trammell et al. (2006)
		Well-being	Garimella and Mihalcea (2016)
		“Conversation” and “At home”	Argamon et al. (2007)
	Men	Persons outside family and peer group	Mazur and Kozarian (2010)
		Society as a whole (e.g. people from their own peer group)	Mazur and Kozarian (2010); Liu and Mihalcea (2007)
Blogging Topics	Women	Homeschooling	Herring et al. (2005b)
		Food and eating	Liu and Mihalcea (2007)
		Creative work	Pedersen and Macafee (2007)
		Fun	Argamon et al. (2007)
		Focus on immediate time, nuanced colors, objects describable by size	Liu and Mihalcea (2007)
	Men	Religion	Argamon, et al. (2007); Campbell (2010), Garimella and Mihalcea (2016)
		Money	Schler et al. (2006)
		Business	Argamon et al. (2007); Miller et al. (2011)
		Internet and technology	Argamon et al. (2007); Pedersen and Macafee (2007); Schler et al. (2006); Miller et al. (2011)
		Science and engineering	Garimella and Mihalcea (2016)
		Links to Internet sites	Pedersen and Macafee (2007)
		Sport	Garimella and Mihalcea (2016)
		Hobbies or interests	Trammell et al. (2006)
		Politics	Argamon et al. (2007); Herring et al. (2005b); Pedersen and Macafee (2007); Liu and Mihalcea (2007); Schler et al. (2006)
		Work-related, cultural	Liu and Mihalcea (2007); Garimella and Mihalcea (2016)
Political Blogging	Women	Political statements, reflection on speeches, record of the day, interviews.	Sweetser (2007)
	Men	Candidate endorsements, bloggers’ credentials, mainstream media, reader interaction, opponent opinions	Sweetser (2007)
Sexua-lity	Men	Homosexual identity	Huffaker and Calvert (2005); Huffaker (2004)
Medical Context	Women	Benefits of psychotherapy and self-help; Resistance to medications; Criticism of doctors	Clarke and Van Amerom (2008)
		Mental health	Miller (2015)
		Obsessive-compulsive disorder (OCD)	Campbell and Longhurst (2013)
		Disease and disability	Miller et al. (2011)
	Men	HIV/AIDS	Miller et al. (2011)
		Provider experiences, health research/news, and health policy	Miller et al. (2011)
		Pharmaceuticals as means to maintain sense of normalcy; Self-harm and suicide; taking action against depression	Clarke and Van Amerom (2008)
Emo-tions	Women	Pleasurable, happy, sad	Liu and Mihalcea (2007)
	Men	Arousal, anger, disgust	Liu and Mihalcea (2007)

Table 5. Gender-relevant findings on blogging content (Findings that show no difference were not included into the table for space reasons, but are available from the authors upon request).

Summarized in Table 5, evidence collected in the blogging context is in some conflict with the behavior observed offline. On the one hand, a stereotypical perspective can be partially confirmed. Indeed, the emotional loading of blogs appears to coincide with common views on female and male self-expression. While females choose pleasurable, happy, or sad topics, men blog on topics provoking anger, arousal or disgust (Liu and Mihalcea, 2007). Personal content, thoughts and concerns are also common topics for female bloggers (Trammell et al., 2006). Further, women like to report on their socialization activities, relationships with close friends, family and a romantic partner (e.g. Mazur and Kozarian, 2010). Interestingly, young men also write about personal relationships: Huffaker and Calvert (2005) show that female and male adolescents equally discuss “boyfriends, girlfriends, or other people they ‘like’[…]” All in all, however, the relationship theme appears to be more dominant for females. Women also engage with such “female” topics as homeschooling (Herring et al., 2005b), or food and eating (Liu and Mihalcea, 2007). Men, on the other hand, adopt a more global view (Mazur and Kozarian, 2010).

Men’s theme coverage is broader, and includes, among others, money (Schler et al., 2006), business (e.g. Argamon et al., 2007), and technology (e.g. Pedersen and Macafee, 2007). Politics, in particular, is popular, providing evidence that male bloggers strive to have an impact beyond their local social networks and a family unit (Argamon et al., 2007). Nonetheless, political engagement of women in blogs is also present. We witness a rise of female blogging dedicated to fighting anti-social behavior (Simmons, 2008), promoting social change (Somolu, 2007) and challenging prevalent gender representations within our society (Lopez, 2009). However, female “social power is significantly weaker than [that of] men in the blogosphere” (Wei, 2009, p. 550). Possible reasons include women’s lack of interest in politics, lack of female quality blogs, and unwillingness of top bloggers to link to female blogs (Harp and Tremayne, 2006).

Despite similarities to offline communication and behavior with regard to content shared when blogging, our analysis suggests that especially men may use blogging in special ways, compensating for the constraints imposed on them offline by e.g., not being encouraged in early childhood to talk about emotions more often (Cross and Madson, 1997). For example, Clarke and Van Amerom (2008)’s examination of blogs by depression patients showed that both men and women were using blogs to emotionally unload themselves. Similar findings were reported for the HIV/AIDS realm, with more male users talking about this diagnosis in blogs (Miller et al., 2011). Furthermore, blogging was frequently used by male homosexual users who perceived this online environment as a safe haven to express their sexual identity and related feelings (Huffaker and Calvert, 2005). All in all, these findings suggest that blogging has a significant potential to provide male users with better access to social support and as a result improve their well-being (Turner, 1981).

4.5. Readership

Differences in motivation to blog may impact the topics chosen by female and male bloggers. Male bloggers prefer a broader perspective and topics with higher appeal to external audience (Mazur and Kozarian, 2010). Women, on the other hand, tend to choose personal themes, concentrating on their own private lives (Pedersen and Macafee, 2007). All readership-relevant insights in the examined articles suggest a higher importance of male voice in the blogosphere, typically measured by the number of incoming links and user traffic (see Table 6). Indeed, our review shows that A-bloggers are predominantly male (Pedersen and Macafee, 2007; Trammell and Keshelashvili, 2005), with men dominating highly ranked political blogs (Harp and Tremayne, 2006). Male blogs also receive more links (Herring et al., 2004b) and have higher readership (Harp and Tremayne, 2006). Moreover, when blogging landscape and bloggers’ opinions are discussed in the mass media, male bloggers are mentioned earlier, more often, and are more likely to be mentioned by name (Herring et al., 2004b). Unattractive topic choice, however, is only one reason behind this apparent lack of popularity of female blogging. For example, Harp and Tremayne (2006) report that even for subject-oriented filter blogs, male blogs are still likely to receive more links and have more readers. Gregg (2006) attributes this to the early advantage men hold: Historically occupying stronger positions in society men find it easier to establish themselves as authoritative and important. Social views on men and women also differ, with masculine qualities being more valued especially in the Western society (Wajcman, 1991). Herring et al. (2004b) suggest that "[…] discourse practices that construct weblogs as externally-focused, substantive, intellectual, authoritative, and potent […] map readily on to Western cultural notions of white collar masculinity". From a more modern perspective Harp and Tremayne (2006) argue that the traditional linking hierarchy of the blogging environment favors men for reasons related to growth and preferential attachment. Since mainly men were early adopters of the Web they have a first-mover advantage which women still need to outrun. A more plausible explanation, however, refers to the engrained power patterns which may still hold their grip, consciously or unconsciously intervening with the behavior of users when it comes to hyperlinking (Gregg, 2006).

Focus	Insights	Source
Men	Blogs of male owners are perceived as more credible.	Armstrong and McAdams (2009)
	British male bloggers more visible.	Pedersen and Macafee (2007)
	The only two British bloggers in the A-list are male.	Pedersen and Macafee (2007)
	More A-list blogs operated by men.	Trammell and Keshelashvili (2005)
	More men operate highly ranked political blogs.	Harp and Tremayne (2006)
	Male bloggers have more incoming links and readers even when only filter blog are considered.	Harp and Tremayne (2006)
	Male blogs receive more attention by other bloggers.	Wei (2009)
	Males are more likely to operate filter blogs, which are more popular with readers.	Herring et al. (2004b)
	More male blogs mentioned in mass media reports.	Herring et al. (2004b)
	The linking behavior favors the blogs of males.	Herring et al. (2004b)

Table 6. Gender-relevant findings on readership in blogging.

4.6. Writing Style

In the offline environment, women and men express themselves in different ways both verbally and nonverbally, resulting in different communication styles. In general, the verbal language used by females is perceived to be more pleasant, polite and personal (Mulac and Lundell, 1980; Tannen, 1994; Haas, 1979). Men, on the other hand, express themselves in a far more direct and factual way (Tannen, 1994; Haas, 1979). This is also in line with the theory of Baumeister and Sommer (1997), according to which men try to achieve a higher position within existing hierarchies by demonstrating their knowledge to others; whereas women try to exhibit close relationships through intimacy. Gender also plays a vital role when it comes to the interpretation of nonverbal expressions. For example, eye contact is perceived as a friendly attitude for women. Yet, the same expression can be perceived as an attempt to dominate for men (Levine and Feldman, 2002). This suggests that nonverbal communication is more important for and to women because in everyday life they are more encouraged to engage in it. Indeed, women are consistently shown to be more expressive about their emotions in nonverbal communication (McClure, 2000; Burgoon, 1994), which, for example, is reflected in their smiling habits (LaFrance et al., 2003). Research evidence suggests that women may be more conscious regarding their choice of words and invest more effort into their verbal and nonverbal communication (Levine and Feldman, 2002; McClure, 2000; Burgoon, 1994).

In the blogging environment (see Table 7), traditional gender roles are mainly reflected in the choice and use of words. Female bloggers are shown to prefer a more personal writing style (Schler at al., 2006) using personal pronouns more, while men use impersonal pronouns (e.g. Argamon et al., 2007; Herring and Paolillo, 2006) and have a rather fact-based writing style (Gill et al., 2005; Van Doorn et al., 2007) with more nouns and prepositions (e.g. Argamon et al., 2007; Nowson and Oberlander, 2006). For example, in the case of health-oriented blogs men write from a more professional perspective whereas women blog from the perspective of patients and caregivers (Miller et al., 2011). This writing style could be a reflection of the types of blogs men and women choose: men favor filter blogs with a rather information-based content and women prefer diary blogs containing personal insights.

Domain	Gender	Insights	Source
Verbal: Special Words	Women	Make-up, jewellery, fabulous, Barbie, layed, kissme, muah (kissing sound)	Yan and Yan (2006)
		Use of purple, tan and pink	Liu and Mihalcea (2007)
		Use of sensorial words	Garimella and Mihalcea (2016)
		Adverbs ending with "ly", evaluative adjectives and intensifiers	Gukosyants (2015)
		Phrases with content words	Nowson and Oberlander (2006)
	Men	Psst, nba, poet, income, badass, furious, wasup	Yan and Yan (2006)
		Use of navy, gold, silver	Liu and Mihalcea (2007)
		Quantitative words	Gukosyants (2015)
		Function word combinations (e.g. "of the")	Nowson and Oberlander (2006)
Verbal: Word Use	Women	Use more personal pronouns (First-person singular, first person plural, third-person masculine)	Argamon et al. (2007); Herring and Paolillo (2006); Liu and Mihalcea (2007); Nowson and Oberlander (2006); Schler et al. (2006)
		Gerunds, metaphors and ellipsis	Gukosyants (2015)
		Use more verbs	Argamon et al. (2007); Nowson and Oberlander (2006)
	Men	Use fewer and more impersonal pronouns (First-person plural, second- and third-person pronouns)	Herring and Paolillo (2006); Liu and Mihalcea (2007); Nowson and Oberlander (2006)
		Seldom use of figural expressions	Gukosyants (2015)
		More articles, nouns and prepositions	Argamon et al. (2007); Nowson and Oberlander (2006); Schler et al. (2006)
Non-verbal: Images	Women	Prefer more images than text	Hsu (2012)
	Men	Prefer more text than images	Hsu (2012)
Non-verbal: Emoticons	Men	Use of more emoticons	Huffaker (2004); Huffaker and Calvert (2005)
General Tone	Women	Personal and involved writing style	Schler at al.(2005)
		More swearwords	Argamon et al. (2007)
		Passive voice	Gukosyants (2015)
		Apologetic and polite words	Gukosyants (2015)
		Immediate writing style, words from the area of sociality, emotions and physical states	Nowson and Oberlander (2006)
		In academic writing - contextual style, women can display formal style no less than men	Gill et al. (2005)
	Men	Formal writing style	Gill et al. (2005); Van Doorn et al. (2007)
		More swearwords	Gukosyants (2015)
		Imperative, rhetorical questions and cohesions	Gukosyants (2015)
		Active and resolute language	Huffaker (2004); Huffaker and Calvert (2005); Gukosyants (2015)
	Both	No differences: formality or tone	Fullwood et al. (2009)
		No differences: use of cooperative and accommodating language	Huffaker (2004); Huffaker and Calvert (2005)
		No differences in use of relative-time compared to concrete-time expression and vice versa	Liu and Mihalcea (2007)

Table 7. Gender-relevant findings on writing style.

Surprisingly, women have been found to use more swear words in blogs (Argamon et al., 2007). This is contrary to the traditional view which implies men to be more aggressive (Cross and Madson, 1997). It appears that some women use the blogosphere to step out of the traditional gender role and behave more freely. Further, there are some indications that the writing style of female bloggers grows "male" with age (Argamon et al., 2007; Schler et al., 2006), suggesting that mature women lean towards a more professional and fact-based style of self-expression.

When it comes to nonverbal communication bloggers use images and emoticons in written text to replace missing facial expressions (Crystal, 2001). While women prefer an aesthetic layout with several images, men prefer a layout with a lot of text (Hsu, 2012), supporting the predominantly emotional concerns of women and the rather information-dominant interests of men (Cross and Madson, 1997). Typically, the use of emoticons is associated with the perception of an author being more open and friendly (Constantin et al., 2002). While findings on the use of emoticons in blogs are scarce, some studies do not find a difference in the use of emoticons (Fullwood et al., 2009), and other authors even report that it is mainly male teenagers who use them (Huffaker, 2004; Huffaker and Calvert, 2005). This, however, does not correspond with a traditional view on male reticence (Tannen, 1994; Cross and Madson, 1997).

All in all, our findings in this area suggest that both men and women use blogging to transgress their traditional gender boundaries, be it in the use of swear words for women (Argamon et al., 2007), or emoticons for men (Huffaker, 2004).

4.7. Privacy

In both offline and online domains women have consistently reported higher privacy concerns. These attitudes can be partially attributed to the higher risk aversion of females (Byrnes et al., 1999). High privacy concerns of women are, however, somewhat in conflict with their communication propensity, needs and skills (Thelwall, 2011). Indeed, women are known to be more intimate and talk more about themselves (Davidson and Duberman, 1982; Jourard, 1971). Together, these patterns suggest complex mechanisms of female self-expression and audience control. For example, in the context of social web, women are consistently shown to self-disclose more, but also be more cautious when doing so (Thelwall, 2011). For example, on Social Networking Sites women are more likely to post pictures (Hoy and Milne, 2010), yet they are less likely to disclose their basic and contact information (Li-Barber, 2012), and they also rely more on privacy settings (e.g. Lewis et al., 2008). Together this logic underlies the Social Web Gendered Privacy Model (SWGPM) advanced by Thelwall (2011). Blogging context offers an interesting environment to explore these dynamics.

Summarized in Table 8, our findings provide partial support for SWGPM. Similar to other contexts, female bloggers appear to be highly concerned about privacy (Pedersen and Macafee, 2007). For privacy behavior, Viégas (2005) differentiates between four areas relevant to blogging: (1) content, (2) identity management, (3) audience and (4) persistence. In terms of (1) content, behavior of female bloggers is generally in line with SWGPM, as they disclose more and are also more intimate in their self-disclosures (Hollenbaugh and Everett, 2013). For example, Bortree (2005) reports that teenage girls in their sample provide most intimate thoughts, frustrations and disappointments in their blogs. However, evidence on (2) the identity management is limited, and is somewhat conflicting with the idea that women should be more cautious. Huffaker and Calvert (2005) document that a comparable number of male and female bloggers have revealed their first name (36% vs. 34%), birth date (21% vs. 18%), age (37% vs. 30%), and contact information (27% vs. 34%). The only difference was information on location that was revealed significantly more by male bloggers (34% vs. 25%). Women, on the other hand, were more likely to provide a URL of their homepage (9% vs. 21%), indirectly undermining their anonymity. Women were more conscious when their anonymity cover was removed: women who used fake or no names disclosed more and on a wider variety of topics than women who used their real names. In contrast, men surprisingly self-disclosed more when they were visually identified. While our review provides findings on content and identity management, the issues of (3) control over the audience and

(4) persistence of information online have not been addressed from the gender perspective, calling for more research in these areas.

Focus		**Findings**	**Source**
Privacy Concerns	Women	Privacy is a major issue.	Pedersen and Macafee (2007)
		More concerned about co-owners of their disclosures.	Child (2007)
Privacy Behaviour: **Content**	Women	Self-disclose more.	Hollenbaugh and Everett (2013); Hollenbaugh (2010)
		More intimate details.	Hollenbaugh and Everett (2013) Li et al. (2015)
		The most divulging bloggers are women who also are generally divulging offline.	Hollenbaugh (2010)
		Disclosure of information increases with age and larger personal social network.	Li et al. (2015)
		Disclose more frequently privacy-relevant information.	Li et al. (2015)
		Women who are less identifiable are more likely to write negative about others.	Fullwood et al. (2013)
		Women who blogged more and shared more private information, are more likely to remove references to past (broken) relationships.	Child et al. (2012)
	Both	Equally likely to disclose information.	Child (2007)
Privacy Behaviour: **Identity Management**	Women	Make a link to personal web site more often.	Huffaker and Calvert (2005)
		Women who used fake or no names disclosed more and on a wider variety of topics than women who used their real names.	Hollenbaugh and Everett (2013)
	Men	More likely to provide information on location.	Huffaker and Calvert (2005)
		More likely to identify themselves.	Fullwood et al. (2013)
		Self-disclosed more information when they were visually identified.	Hollenbaugh and Everett (2013)
	Both	No difference in the disclosure of first / full name, age, birthday, contact info, email, IM.	Huffaker and Calvert (2005)
Past Experience	Both	Both experienced problems resulting from blogging.	Pedersen and Macafee (2007)
Sharing	Women	Trust is a critical factor of knowledge sharing behavior.	Chai et al. (2011)
	Men	Information privacy concerns have a negative influence on knowledge sharing behavior.	Chai et al. (2011)

Table 8. Gender-relevant findings on privacy in blogging.

5. Discussion and Future Research Agenda

Blogging created a disruption in the media world by bringing new and previously unestablished media actors – bloggers - to the front of the new media (Vaast et al., 2013). As such, blogging holds potential to empower individuals to establish their unique voice and presence in the new world of social media. Thus far, blogging has been widely adopted by both male and female bloggers as a tool to share their personal, professional and societal perspectives. However, male voice appears to have a stronger impact in the blogosphere and beyond (e.g. Herring et al., 2004b), raising troubling questions about the existence of gender parity on the social web. To better understand differences in male and female use of blogging we make several contributions to research and practice. First, we structure and summarize gender relevant insights in the blogging domain. Second, we contribute to gender discourse by identifying behavior and communication styles preferred by female and male bloggers and ways in which the blogosphere promotes a shift in gender roles. Finally, as recommended by Webster and Watson (2002), we identify critical knowledge gaps and advocate for future research on gender issues in the blogosphere (Table 9).

Theme	Similarity to traditional offline perspective on gender	Deviations from traditional offline perspective on gender	Research gaps / Suggestions for future research
Motivation to Blog	Women are motivated by social connection, self-expression and documenting; Men by professional use, attention and reputation, and "being an information hub".	No differences in the motivational strength of some determinants, otherwise typical for gender research (e.g. "being an information hub).	Very limited evidence on gender-specific motivational patterns calls for more comprehensive research in this area.
Types of Blogs	Men operate filter, knowledge and mixed blogs; Women prefer personal, diary, journal blogs.	Diary blogs are used by both sexes. Women with high standards of achievement tend to operate filter blogs.	Need for more studies of women operating filter, knowledge and mixed blogs, and men running personal journal blogs – motives, patterns of use and relationship to social support and empowerment.
Bloggers' Characteristics	Men have blogged for longer. Women are more active, interested in their audience, and invest more efforts into blogs.	Men cooperate, mention others and stay in contact via RSS feeds more often than women.	Need to better understand whether and how blogging motivates users to transgress their traditional gender boundaries. For example, do blogs help men to step out of their traditional roles and get more access to social support? Do females use blogs to release stress and aggression? What are the style and content characteristics of successful female blogs? Are female bloggers as competitive as male bloggers?
Blog Content	Women focus on personal and family/ home topics; Men discuss society, technology, politics and money.	Men in previously marginalized groups also focus on personal topics.	
Writing Style	Female bloggers have personal, while male bloggers have more impersonal fact-based writing style.	Young males more likely to use emoticons to express emotions. Women swear more in blogs.	
Readership	Male blogs are more popular among readers, media and other bloggers.	Not available.	Reasoning behind lower popularity of female blogs (e.g. unattractive topic choice, historical advantage of men, engrained power patterns) is hypothetical. Research should investigate the causes of this power dichotomy to help mitigate these biased trends.
Privacy	Female bloggers concerned about privacy; but talk more about personal and intimate topics.	Gender-specific evidence on identity management is limited, and is in conflict with the idea that women are more cautious.	Control over audience and persistence of information – core areas of bloggers' privacy – have not been addressed from the gender perspective, calling for more research.

Table 9. Summary of findings and directions for future research.

Our literature review identifies gender differences and similarities in various aspects of blogging, uncovering both consistency but also some divergence from the traditional view on gender offline (e.g. Cross and Madson, 1997; Tannen, 1994). We observe that while male bloggers were the earliest adopters, female bloggers have gained much ground and by some estimates now even outnumber men in the blogosphere. In line with previously observed offline communication and behavior patterns, male bloggers are goal-oriented and focus on information dissemination and furthering their career. Women, on the other hand, are more relationship-oriented and blog to self-express and connect with others (Table 3). Moreover, the topics men and women choose differ dramatically consistent with traditional gender roles in the society: Male bloggers prefer socially-relevant topics in politics, business, money and technology, while female writing is personal and emotional, mostly focused on relationships, family and domestic issues (Table 5). As a result, male bloggers usually operate filter and knowledge blogs, and women choose diary and journal blogs.

Gender roles in society, however, continue to evolve and more and more males and females are crossing gender boundaries to pursue activities that would have been considered untraditional only some decades ago. In the blogging context some of these changes are also spurred by the specifics of this online medium. For example, a blogger may choose to stay anonymous, which allows him or her to more openly express feelings and opinions with less concern for public backlash or need to conform to stereotypes. Anonymity of blogging appears to relieve social constraints, enabling men, just like women, to share deeply personal thoughts and seek help from others. Women also profit from being able to express their aggressions and frustrations in blogging by swearing – an important way to release stress and enhance well-being. Further, since bloggers rely on the larger online community to read, comment and re-share their contributions, social connection and cooperation emerges as important. Thus, to be successful (which is innately important to goal-oriented men) bloggers need to engage with others and maintain relationships (traditionally female activity). Thus, findings on cooperative tendencies of male bloggers may be partially attributed to their shifting preferences for connectedness and their need to self-promote and establish themselves.

Our findings also shed light on the empowering impact of blogs for female bloggers. Indeed, much has been done to encourage women to become equitable members of the modern society in all spheres of life to achieve parity in the workplace and especially in leadership. In this context, our research suggests that blogging may work as an empowerment tool that gives women a formidable voice and allow them to be engaged with others in meaningful ways. Yet, much remains to be done to achieve complete parity. Unfortunately, readership of female blogs still lags that of male bloggers. Given the plethora of findings supporting traditional communication and blogging patterns of men and women, we observe that women prefer personal topics that seem to have lower relevance to the greater community as shown by insights related to readership where male dominate the field. When striving for equal representation in media and public sentiment and to capitalize on the economic benefits offered by blogging, women should, among other, try to appeal to a wider audience in their blogs and support one another by linking to other relevant bloggers.

All in all, while much research has been done on the various aspects of blogging, many issues in gender discourse remain unaddressed. Table 9 identifies research gaps across seven themes that we used to structure our literature review. Moreover, most of the studies in our review were based on data collected in North-America using English language blogs, and thus are not globally representative. Hence, since gender is partially socially constructed (Eagly, 1987; Eagly and Steffen, 1984), exploration of international blogs, especially representing non-western cultures and developing world, presents a major opportunity for future research.

6. Limitations

Our study has several limitations. First, current literature typically defines blogging success and impact based on the number of followers, readers and shares. However, perceptions and measures of blogging success may differ based on gender. For example, while some bloggers may be concerned with audience size, others may be more interested in audience engagement and support. As a result existing research findings showing stronger impact of male bloggers may be biased because the success measures of audience quantity are much easier to capture than audience quality. Second, our sample included only studies in English leading to possible cultural bias in our findings since many of these studies use North American data. Finally, we were only able to review publicly available studies which could lead to publication bias.

References

List of studies included in the review

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2007). Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, *12*(9).

Armstrong, C. L., & McAdams, M. J. (2009). Blogs of information: How gender cues and individual motivations influence perceptions of credibility. *Journal of Computer-Mediated Communication*, *14*(3), 435-456.

Blinka, L., & Smahel, D. (2009). Fourteen is fourteen and a girl is a girl: Validating the identity of adolescent bloggers. *Cyberpsychology & Behavior*, *12*(6), 735-739.

Bortree, D. S. (2005). Presentation of self on the web: An ethnographic study of teenage girls' weblogs. *Education, Communication & Information*, *5*(1), 25-39.

Campbell, H. A. (2010). Religious authority and the blogosphere. *Journal of Computer-Mediated Communication*, *15*(2), 251-276.

Campbell, R., & Longhurst, R. (2013). Obsessive–compulsive disorder (OCD): Gendered metaphors, blogs and online forums. *New Zealand Geographer*, *69*(2), 83-93.

Carstensen, T. (2009). Gender Trouble in Web 2.0. Gender perspectives on social network sites, wikis and weblogs. *International Journal of Gender, Science and Technology*, *1*(1).

Chai, S., Das, S., & Rao, H. R. (2011). Factors affecting bloggers' knowledge sharing: An investigation across gender. *Journal of Management Information Systems*, *28*(3), 309-342.

Chen, G. M. (2012). Why do women write personal blogs? Satisfying needs for self-disclosure and affiliation tell part of the story. *Computers in Human Behavior*, *28*(1), 171-180.

Child, J. T. (2007). *The development and test of a measure of young adult blogging behaviors, communication, and privacy management*. ProQuest.

Child, J. T., Haridakis, P. M., & Petronio, S. (2012). Blogging privacy rule orientations, privacy management, and content deletion practices: The variability of online privacy management activity at different stages of social media use. *Computers in Human Behavior*, *28*(5), 1859-1872.

Clarke, J., & Van Amerom, G. (2008). A comparison of blogs by depressed men and women. *Issues in Mental Health Nursing*, *29*(3), 243-264.

Fullwood, C., Sheehan, N., & Nicholls, W. (2009). Blog function revisited: A content analysis of MySpace blogs. *CyberPsychology & Behavior*, *12*(6), 685-689.

Fullwood, C., Melrose, K., Morris, N., & Floyd, S. (2013). Sex, blogs, and baring your soul: factors influencing UK blogging strategies. *Journal of the Association for Information Science and Technology*, *64*(2), 345-355.

Fullwood, C., Nicholls, W., & Makichi, R. (2015). We've got something for everyone: How individual differences predict different blogging motivations. *New Media & Society*, *17*(9), 1583-1600.

Gill, A. J., Nowson, S., & Oberlander, J. (2005). Weblogs, Genres, and Individual Differences. In *Proceedings of the Cognitive Science Society* (Vol. 27, No. 27).

Garimella, A., & Mihalcea, R. (2016). Zooming in on gender differences in social media. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, Osaka, Japan (pp. 1–10).

Gregg, M. C. (2006). *Posting with passion: Blogs and the politics of gender*. In Bruns, A. & Jacobs, J. (Eds.). *Uses of blogs*. New York, NY: Peter Lang, 151–160.

Guadagno, R. E., Okdie, B. M., & Eno, C. A. (2008). Who blogs? Personality predictors of blogging. *Computers in Human Behavior*, *24*(5), 1993-2004.

Gukosyants, O. (2015). Gender Identity and Language Behaviour of English-Speaking Internet Users (Internet Chat Room and Blog Users). *Russian Review* (1), 2-2.

Harp, D., & Tremayne, M. (2006). The gendered blogosphere: Examining inequality using network and feminist theory. *Journalism & Mass Communication Quarterly*, *83*(2), 247-264.

Herring, S. C., & Paolillo, J. C. (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics*, *10*(4), 439-459.

Herring, S. C., Kouper, I., Paolillo, J. C., Scheidt, L. A., Tyworth, M., Welsch, P. & Yu, N. (2005b). Conversations in the blogosphere: An analysis" from the bottom up". In *HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on System Sciences* (pp. 107b-107b). IEEE.

Herring, S. C., Kouper, I., Scheidt, L. A., & Wright, E. L. (2004b). Women and children last: The discursive construction of weblogs. In L. Gurak, S. Antonijevic, L. Johnson, C. Ratliff & J. Reyman (Eds.), Into the Blogosphere: Rhetoric, Community, and Culture of Weblogs.

Herring, S. C., Scheidt, L. A., Bonus, S., & Wright, E. (2004a). Bridging the gap: A genre analysis of weblogs. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*, Los Alamitos: IEEE Press.

Herring, S. C., Scheidt, L. A., Wright, E., & Bonus, S. (2005a). Weblogs as a bridging genre. *Information Technology & People*, *18*(2), 142-171.

Hollenbaugh, E. E. (2010). Personal journal bloggers: Profiles of disclosiveness. *Computers in Human Behavior*, *26*(6), 1657-1666.

Hollenbaugh, E. E. (2011). Motives for maintaining personal journal blogs. *Cyberpsychology, Behavior, and Social Networking*, *14*(1-2), 13-20.

Hollenbaugh, E. E., & Everett, M. K. (2013). The effects of anonymity on self-disclosure in blogs: An application of the online disinhibition effect. *Journal of Computer-Mediated Communication*, *18*(3), 283-302.

Hsu, C. C. (2012). Comparison of gender differences in young people's blog interface preferences and designs. *Displays*, *33*(3), 119-128.

Hsu, C. L., and Lin, J. C. C. (2008). Acceptance of blog usage: The roles of technology acceptance, social influence and knowledge sharing motivation. *Information & Management* 4(1), 65-74.

Huang, C. Y., Shen, Y. Z. & Lin, H. X., & Chang, S. S. (2007). Bloggers' motivations and behaviors: A model. *Journal of Advertising Research*, *47*(4), 472-484.

Huffaker, D. (2004). *Gender similarities and differences in online identity and language use among teenage bloggers* (Doctoral dissertation, Georgetown University).

Huffaker, D. A., & Calvert, S. L. (2005). Gender, identity, and language use in teenage blogs. *Journal of Computer-Mediated Communication*, *10*(2).

Kennedy, T. L., Robinson, J. S., & Trammell, K. (2005). Does gender matter? Examining conversations in the blogosphere. *Internet Research*, *6*, 5-9.

Li, Q. (2005). Gender and CMC: A review on conflict and harassment. *Australasian Journal of Educational Technology*, *21*(3), 382-406.

Li, K., Lin, Z., & Wang, X. (2015). An empirical analysis of users' privacy disclosure behaviors on social network sites. *Information & Management*, *52*(7), 882-891.

Liao, H. L., Liu, S. H., & Pi, S. M. (2011). Modeling motivations for blogging: An expectancy theory analysis. *Social Behavior and Personality: an international journal*, *39*(2), 251-264.

Liu, H., & Mihalcea, R. (2007). Of Men, Women, and Computers: Data-Driven Gender Modeling for Improved User Interfaces. *ICWSM* (7), pp. 26-28.

Liu, Y. (2017). Gender Difference in Perception and Use of Social Media Tools. In *Handbook of Research on Instructional Systems and Educational Technology*(pp. 249-262). IGI Global.

Lopez, L. K. (2009). The radical act of 'mommy blogging': redefining motherhood through the blogosphere. *New Media & Society*, *11*(5), 729-747.

Lu, H. P., & Hsiao, K. L. (2009). Gender differences in reasons for frequent blog posting. *Online Information Review*, *33*(1), 135-156.

Lu, H. P., Lin, J. C. C., Hsiao, K. L., & Cheng, L. T. (2010). Information sharing behaviour on blogs in Taiwan: Effects of interactivities and gender differences. *Journal of Information Science*, *36*(3), 401-416.

Mazur, E., & Kozarian, L. (2010). Self-presentation and interaction in blogs of adolescents and young emerging adults. *Journal of Adolescent Research*, *25*(1), 124-144.

Mazur, E., & Li, Y. (2016). Identity and self-presentation on social networking web sites: A comparison of online profiles of Chinese and American emerging adults. *Psychology of Popular Media Culture*, *5*(2), 101-118.

Miller, E. A., Pole, A., & Bateman, C. (2011). Variation in health blog features and elements by gender, occupation, and perspective. *Journal of Health Communication*, *16*(7), 726-749.

Miller, E. A., Pole, A., & Usidame, B. (2015). Life and Death in the Mental-Health Blogosphere: An Analysis of Blog Content and Survival. *World Medical & Health Policy*, *7*(1), 59-82.

Nardi, B. A., Schiano, D. J., Gumbrecht, M., & Swartz, L. (2004). Why we blog. *Communications of the ACM*, *47*(12), 41-46.

Nowson, S., & Oberlander, J. (2006). The Identity of Bloggers: Openness and Gender in Personal Weblogs. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* (pp. 163-167).

Pedersen, S., & Macafee, C. (2007). Gender differences in British blogging. *Journal of Computer-Mediated Communication*, *12*(4), 1472-1492.

Porter, L., Sweetser, K., & Chung, D. (2009). The blogosphere and public relations: Investigating practitioners' roles and blog use. *Journal of Communication Management*, *13*(3), 250-267.

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W. (2006). Effects of Age and Gender on Blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* (Vol. 6, pp. 199-205).

Simmons, T. (2008). The personal is political? Blogging and citizen stories, the case of Mum's Army. *Information Polity*, *13*(1-2), 87-96.

Somolu, O. (2007). 'Telling our own stories': African women blogging for social change. *Gender & Development*, *15*(3), 477-489.

Stavrositu, C., & Sundar, S. S. (2012). Does blogging empower women? Exploring the role of agency and community. *Journal of Computer-Mediated Communication*, *17*(4), 369-386.

Sweetser, K. D. (2007). Blog bias: Reports, inferences, and judgments of credentialed bloggers at the 2004 nominating conventions. *Public Relations Review*, *33*(4), 426-428.
Trammell, K. D., & Keshelashvili, A. (2005). Examining the new influencers: A self-presentation study of A-list blogs. *Journalism & Mass Communication Quarterly*, *82*(4), 968-982.
Trammell, K. D., Tarkowski, A., Hofmokl, J., & Sapp, A. M. (2006). Rzeczpospolita blogów [Republic of Blog]: Examining Polish bloggers through content analysis. *Journal of Computer-Mediated Communication*, *11*(3), 702-722.
Van Doorn, N., Van Zoonen, L., & Wyatt, S. (2007). Writing from Experience Presentations of Gender Identity on Weblogs. *European journal of women's studies*, *14*(2), 143-158.
Viégas, F. B. (2005). Bloggers' expectations of privacy and accountability: An initial survey. *Journal of Computer-Mediated Communication*, *10*(3).
Wei, L. (2009). Filter blogs vs. personal journals: Understanding the knowledge production gap on the Internet. *Journal of Computer-Mediated Communication*, *14*(3), 532-558.
Yan, X., & Yan, L. (2006). Gender Classification of Weblog Authors. In AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs (pp. 228-230).

Other references

Baumeister, R. F., & Sommer, K. L. (1997). What do men want? Gender differences and two spheres of belongingness: comment on Cross and Madson (1997). *Psychological Bulletin*, *122*(1), 38-44.
Bem, S. L. (1981). Gender schema theory: A cognitive account of sex typing. *Psychological review*, *88*(4), 354 - 364.
Broverman, D. M., Klaiber, E. L., & Kobayashi, Y. (1968). Roles of activation and inhibition in sex differences in cognitive abilities. *Psychological Review*, *75*(1), 23-50.
Burgoon, J. K. (1994). *Nonverbal signals*. In Knapp, M. L. & Miller, G. R. (Eds.), *Handbook of Interpersonal Communication*, Thousand Oaks (CA): Sage, 229–85.
Butler, B. S. (2001). Membership size, communication activity, and sustainability: A resource-based model of online social structures. *Information Systems Research*, *12*(4), 346-362.
Byrnes, J. P., Miller, D. C., & Schafer, W. D. (1999). Gender differences in risk taking: A meta-analysis. *Psychological Bulletin*, *125*(3), 367-383.
Cavanaugh, T. (2002). *Let slip the blogs of war*. In Rodzvilla, J. (Ed.), *We've Got Blog: How Weblogs are Changing Our Culture*, Cambridge, MA: Perseus, 188-197.
Chau, M., & Xu, J. (2012). Business intelligence in blogs: Understanding consumer interactions and communities. *MIS Quarterly*, *36*(4), 1189-1216.
Christie, A. A. (2005). How adolescent boys and girls view today's computer culture. *Meridian: A Middle School Computer Technologies Journal*, *8*(1).
Clark, M. L., & Bittle, M. L. (1992). Friendship expectations and the evaluation of present friendships in middle childhood and early adolescence. *Child Study Journal*, 22, 115–135.
Constantin, C., Kalyanaraman, S., Stavrositu, C., & Wagoner, N. (2002). Impression formation effects in moderated chatrooms: An experimental study of gender differences. In *88th Annual Meeting of the National Communication Association, New Orleans, LA*.
Cross, S. E., & Madson, L. (1997). Models of the self: self-construals and gender. *Psychological Bulletin*, *122*(1), 5-37.
Crystal, D. (2001). *Language and the Internet*. Cambridge University Press, Cambridge.
Davidson, L. R., & Duberman, L. (1982). Friendship: Communication and interactional patterns in same-sex dyads. *Sex Roles*, *8*(8), 809-822.
Deci, E. L. (1975). *Intrinsic motivation*. New York, NY: Plenum Press.
Eagly, A. H. (1987). *Sex differences in social behavior: A social-role interpretation*. Lawrence Erlbaum Associates Inc., NJ: Hillsdale.
Eagly, A. H., & Steffen, V. J. (1984). Gender stereotypes stem from the distribution of women and men into social roles. *Journal of personality and social psychology*, *46*(4), 735-754.
Furman, W., & Buhrmester, D. (1985). Children's perceptions of the personal relationships in their social networks. *Developmental psychology*, *21*(6), 1016-1024.
Gefen, D., & Ridings, C. M. (2005). If you spoke as she does, sir, instead of the way you do: a sociolinguistics perspective of gender differences in virtual communities. *ACM SIGMIS Database*, *36*(2), 78-92.
Gefen, D., & Straub, D. W. (1997). Gender differences in the perception and use of e-mail: An extension to the technology acceptance model. *MIS Quarterly*, 389-400.
Haas, A. (1979). Male and female spoken language differences: Stereotypes and evidence. *Psychological Bulletin*, *86*(3), 616-626.
Hafkin, N. J., & Huyer, S. (2007). Women and Gender in ICT Statistics and Indicators for Development. *Information Technologies & International Development*, *4*(2), 25-41.

Hilbert, M. (2011). Digital gender divide or technologically empowered women in developing countries? A typical case of lies, damned lies, and statistics. In *Women's Studies International Forum* (Vol. 34, No. 6, pp. 479-489). Pergamon.
Hoy, M. G., & Milne, G. (2010). Gender differences in privacy-related measures for young adult Facebook users. *Journal of Interactive Advertising*, *10*(2), 28-45.
Jackson, L. A., Ervin, K. S., Gardner, P. D., & Schmitt, N. (2001). Gender and the Internet: Women communicating and men searching. *Sex Roles*, *44*(5), 363-379.
Joiner, R., Gavin, J., Duffield, J., Brosnan, M., Crook, C., Durndell, A., Maras, P., Miller, J., Scott, A. J. & Lovatt, P. (2005). Gender, Internet identification, and Internet anxiety: Correlates of Internet use. *CyberPsychology & Behavior*, *8*(4), 371-378.
Jourard, S. M. (197l). *The transparent self*. Princeton, NJ: Van Nostrand.
Kelleher, T., & Miller, B. M. (2006). Organizational blogs and the human voice: Relational strategies and relational outcomes. *Journal of Computer‐Mediated Communication*, *11*(2), 395-414.
Kim, H. W., Chan, H. C., & Gupta, S. (2007). Value-based adoption of mobile internet: an empirical investigation. *Decision Support Systems*, *43*(1), 111-126.
LaFrance, M., Hecht, M. A., & Paluck, E. L. (2003). The contingent smile: a meta-analysis of sex differences in smiling. *Psychological Bulletin*, *129*(2), 305–334.
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.
Levine, S. P., & Feldman, R. S. (2002). Women and men’s nonverbal behavior and self-monitoring in a job interview setting. *Applied HRM Research*, *7*(1), 1-14.
Levy, Y., & Ellis, T. J. (2006). A systems approach to conduct an effective literature review in support of information systems research. *Informing Science*, *9*.
Lewis, K., Kaufman, J., & Christakis, N. (2008). The taste for privacy: An analysis of college student privacy settings in an online social network. *Journal of Computer‐Mediated Communication*, *14*(1), 79-100.
Li-Barber, K. T. (2012). Self-disclosure and student satisfaction with Facebook. *Computers in Human Behavior*, *28*(2), 624-630.
McClure, E. B. (2000). A meta-analytic review of sex differences in facial expression processing and their development in infants, children, and adolescents. *Psychological Bulletin*, *126*(3), 424–453.
Meyers-Levy, J. (1989). *Gender Differences in Information Processing: A Selectivity Interpretation*. In Cafferata, P. & and Tybout, A. (Eds.), *Cognitive and Affective Responses to Advertising*, Lexington, MA: Lexington, 219-260.
Mulac, A., & Lundell, T. L. (1980). Differences in perceptions created by syntactic‐semantic productions of male and female speakers. *Communications Monographs*, *47*(2), 111-118.
Nielsen (2012). Buzz in the Blogosphere: Millions More Bloggers and Blog Readers. Retrieved from http://www.nielsen.com/us/en/newswire/2012/buzz-in-the-blogosphere-millions-more-bloggers-and-blog-readers.html.
Pingdom (2013). Blog readership demographics – investigating the world’s top blogs. Retrieved from http://royal.pingdom.com/2013/03/01/blog-readership-demographics-2013/.
Sifry, D. (2004). Sifry’s Alerts: October 2004 State of the Blogosphere. Retrieved from http://www.sifry.com/alerts/archives/000390.html
Stewart-Williams, S., & Thomas, A. G. (2013). The ape that thought it was a peacock: Does evolutionary psychology exaggerate human sex differences?. *Psychological Inquiry*, *24*(3), 137-168.
Sysomos (2010). Inside Blog Demographics. Retrieved from https://www.sysomos.com/reports/bloggers/.
Tannen, D. (1994). *Gender and discourse*. New York, NY: Oxford University Press.
TechnoratiMedia (2013). Digital Influence Report 2013. Retrieved from http://www.slideshare.net/truthdefender/technorati-2013-digital-influence-report.
Teo, T. S., & Lim, V. K. (2000). Gender differences in internet usage and task preferences. *Behaviour & Information Technology*, *19*(4), 283-295.
Thelwall, M. (2011). *Privacy and gender in the Social Web*. In Reinecke, L. (Ed.), *Privacy online: Perspectives on Privacy and Self-Disclosure in the Social Web*. New York, NY: Springer, 255-269.
Turner, R. J. (1981). Social support as a contingency in psychological well-being. *Journal of Health and Social Behavior*, 357-367.
Vaast, E., Davidson, E. J., & Mattson, T. (2013). Talking about Technology: The Emergence of a New Actor Category Through New Media. *Mis Quarterly*, *37*(4), 1069-1092.
Venkatesh, V., & Morris, M. G. (2000). Why don't men ever stop to ask for directions? Gender, social influence, and their role in technology acceptance and usage behavior. *MIS Quarterly*, 115-139.
Wajcman, J. (1991). *Feminism confronts technology*. University Park, PA: Pennsylvania State University Press.
Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, 26(2), 13-23.

ARTICLE 12:

MEN, WOMEN, MICROBLOGGING WHERE DO WE STAND?

Authors: Annika Baumann, Humboldt-Universität zu Berlin, Berlin, Germany

Hanna Krasnova, Universität Potsdam, Potsdam, Germany

Natasha Veltri, University of Tampa, Tampa, Florida, USA

Yunsi Ye, Humboldt-Universität zu Berlin, Berlin, Germany

Proceedings: 12. Internationalen Tagung Wirtschaftsinformatik (WI 2015)

Publication year: 2015

Abstract

With millions of users worldwide, microblogging has developed into a powerful tool for interaction and information dissemination. While both men and women readily use this technology, there are significant differences in how they embrace it. Understanding these differences is important to ensure gender parity, provide advertisers with actionable insights on the marketing potential of both groups, and to inform current theories on how microblogging affordances shape gender roles. So far, existing research has not provided a unified frame-work for such analysis, with gender insights scattered across multiple studies. To fill this gap, our study conducts a comprehensive meta-review of existing re-search. We find that current discourse offers a solid body of knowledge on gen-der differences in adoption, shared content, stylistic presentation, and a rather convoluted picture of female and male interaction. Together, our structured findings offer a deeper insight into the underlying dynamics of gender differences in microblogging.

Keywords: *Microblogging, Twitter, Gender Differences, Meta-Review.*

1. Introduction

Having imposed its famous “140-character limit” in 2006, Twitter has coined the term microblogging (MB). Since then it remains the most widely used MB platform with over 255 million monthly active accounts and is ranked as the fifth most popular social network in the world (Statista, n.d.). Spurred by its success, other MB platforms have mushroomed, including such popular ones as Tumblr and Sino Weibo in China.

Originally designed to broadcast and share information about user’s activities and opinions in an easy format (Java et al., 2007), MB was quickly embraced by the global online community as a tool for fast dissemination of the most current information. Now companies, politicians and celebrities increasingly rely on MB to stay connected with their audiences and promote their views. However, as MB continues to evolve and become more powerful, a growing number of stakeholders question whether MB equitably benefits different population groups, including minorities, females, and users from remote geographic regions (e.g. Pal and Counts, 2011; Yang et al, 2013). Particularly, gender differences have been discussed in the MB research for a number of social, theoretical and practical reasons. Specifically, from a social perspective, understanding the differences in male and female use of MB for information dissemination is essential to ensure gen-der parity and a globally equitable information society (Hafkin, 2008). From a theoretical stand-point, past research has identified significant gender differences across a variety of IT contexts, suggesting that certain functional affordance may support, but also change traditional gender roles and behaviors (e.g. Huffaker and Calvert, 2005; Venkatesh and Morris, 2000). Considering the popularity of MB, it is hence critical to understand the role that MB may play in these processes. Finally, understanding gender differences in MB is important for practical reasons: MB has a great potential to viral spreading, e.g., in form of candy or firestorms. Hence, understanding gender specific behavior - for example by identifying user groups that are more likely to reshare, or the presence of gender homophily in resharing – can be an important step in managing word-of-mouth in MB channels (Chu and Kim, 2011).

Reflecting importance of this topic for both theory and practice, many studies focus directly on gender differences in MB (e.g., Bamman et al., 2012; Herdağdelen and Baroni, 2011) or provide supplementary gender-related findings (e.g., Bontcheva et al., 2013; Case and King, 2012). However, existing insights remain scattered, calling for a more systematic analysis. To fill this gap, this study conducts a comprehensive meta-review of existing research to provide a deeper insight into gender differences and similarities, and to give impetus to future research efforts in this area.

2. Theoretical Background

2.1. Microblogging

A specific set of core functional features is definitional for any MB platform: Participants can subscribe ("follow") each other via unidirectional relationships to get new status updates in the form of short messages (“tweets”) from users they follow – their “followees”. These messages can further be distributed in the network by re-posting (“retweeting”) them. Additionally, some MB platforms offer extended functionality to their users. For example, Sina Weibo allows for broader range of publish-able media, additional settings for profiles, and a threaded comment system stimulating social interactions (Chen et al., 2011). Furthermore, while Sina Weibo and Twitter still apply a “140-characher” limit, Tumblr does not impose such limitation.

Attracted by the success of MB, research in this domain has been on the rise, with scholars initially focusing on the specifics of user behavior on these platforms. Examples include research on motivations to use MB (e.g. Pentina et al., 2014), privacy considerations (e.g. Humphreys et al., 2010), continued use (Pentina et al., 2014; Wakefield et al., 2011) and approaches to user profiling (e.g. Alowibdi et al., 2013a; Alowibdi et al., 2013b; Burger et al., 2011). As the use of MB matures, particular uses in specific contexts become apparent attracting further scholarly attention. Examples include the use of MB in politics (e.g. Crigler and Just, 2013; Jackson and Lilleker, 2011) and in disaster management (e.g. Lachlan et al., 2014; Mandel et al., 2012). Recognizing an important role of gender discourse in the MB domain these studies often document gender differences as part of their supplementary findings (Bontcheva et al., 2013; Case and King, 2012; Guan et al., 2014) or even address gender differences as a key focus of their studies e.g. Bamman et al., 2012; Herdağdelen and Baroni, 2011; Walton and Rice, 2013). Nonetheless, these gender-related in-sights remain disorganized, impeding research progress in this area.

2.2 Gender Differences in Offline and Technology Settings

The evolutionary psychology explains gender differences by human ancestral past which over the time fostered women to be more compassionate and men more competitive (Stewart-Williams and Thomas, 2013). As a result, women are better prepared for family-life routines and men have better skills to advance in their careers. The developmental social psychology offers a complementary view and asserts that gender differences develop in response to societal role expectations (Eagly, 1987). From the early age boys are encouraged to be independent, competitive and assertive, while girls are expected to be modest, social and nurturing (Jackson et al., 2001). Consequently, men learn to define themselves in terms of their separatedness from others, drawing their self-esteem from the level of their autonomy (Cross and Madson, 1997). Women, in contrast, have a more pronounced tendency to define themselves in terms of their connectedness to others. They focus on relationships, cooperation; seek closeness and emotional support (Cross and Madson, 1997). These particularities naturally impact the way men and women use and benefit from information systems. When it comes to technology adoption existing research finds men to be rather driven by their attitudes to-wards a new application, while women are influenced by opinions of others and relational uses of IT (Venkatesh and Morris, 2000). In the online context, men are shown to spend more time re-searching on the Web, as they strive to inform themselves or solve certain tasks (Gefen and Ridings, 2005). Women, in contrast, use the email more (Jackson et al., 2001). Observing these differences it is natural to expect that the varied innate characteristics and social roles are also likely to trans-late into distinct gender patterns when it comes to MB.

3. Methodology

We followed recommendations of Webster and Watson (2002) and Levy and Ellis (2006) for our meta-review. Initially, we performed a keyword search in various databases (ScienceDirect, EBSCOhost, Wiley Online Library, ACM Digital Library, IEEE Xplore Digital Library, Taylor & Francis Online, JSTOR, Google Scholar) considering only English language sources. Combinations of the following keywords were applied: {microblogs, microblogging, Twit-ter, Weibo} and {gender differences, gender, female, male, woman, women, man, men, girl, boy}. Only studies related to Twitter and Sina Weibo (later referred to as "Weibo") were considered, since these are the most popular MB services using a 140-character limit. We did not limit the search by the year of publication. Articles identified as potentially relevant were scanned using the in-text search for relevant markers (e.g. "female", "male"). We mainly concentrated on academic work that explicitly addressed specific differences between men and women with regard MB, whereas articles that dealt with only female (or male) perspective were excluded from our research. The second step involved a backward and forward search.

The final sample included 60 studies published between 2009 and 2014. Of those, 48 focused on Twitter, eleven on Weibo and one on both. 45% of articles appeared in journals, and 48% in conferences. 16 articles specifically focused on gender aspects. Most often real data extracted from microblogs was used (66.7%) and then content analyzed. Next, all articles were reviewed and relevant findings were extracted and organized into smaller chunks of information – *insights*. In total 205 insights were identified. Two authors have independently reviewed the resultant material to identify a set of leading themes which were then compared and discussed. This approach al-lowed us to uncover five dominant themes (Table 1). Finally, each insight was as-signed to a specific theme by two coders. Inter-coder reliability measured by Cohen's kappa reached 0.808, providing evidence of a high level of agreement (Landis and Koch, 1977). The final decision on the assignment of items to themes was reached by consensus.

Theme	**Theme Description –Gender differences in:**	**Share**
Adoption	…the use of MB and posting frequency.	16.67%
Content	…the choice of the microblog topic and specific content.	16.67%
Audience	…the interaction of users and their perceptions of it.	25.00%
Motivation	…motivational patterns of microbloggers.	4.90%
Presentation	…the writing style, layout, sentiment and word choice.	36.76%

Table 1. Themes in gender-relevant discourse on microblogging

4. Results: Gender Differences in Microblogging

4.1. Gender Differences in the Adoption of Microblogs

Men appear to have used Twitter for longer, suggesting more men among early adopters (Lasorsa, 2012). Currently an overrepresentation of females is reported by web analytics platforms (e.g. Alexa, n.d.).

	Females are more likely to:	**Males are more likely to:**
Usage	– use MB (Cheng et al., 2009; Dar and Shah, 2013; Fox et al., 2009; Hargittai and Litt, 2011; Heil and Piskorski, 2009; Lenhart et al., 2010; Tufekci and Wilson, 2012; Walton and Rice, 2013; *Zhang et al., 2013*[16])	– use MB (Bontcheva et al., 2013; *Fu and Chau, 2013*; Mislove et al., 2011)
	– *managers/politicians are equally likely to use MB* (Vergeer and Hermans, 2013; Wamba and Carter, 2013)	
Frequency	– post on MB (Burger et al., 2011; Herdağdelen, 2013; Jackson and Lilleker, 2011; Soedjono, 2012; *Zhang et al., 2013*)	– post on MB (Bontcheva et al., 2013; *Fu and Chau, 2013*; Lasorsa, 2012)
	– *no difference in frequency of posting* (Heil and Piskorski, 2009; Rao et al., 2010)	
Specifics of Use	– be active: midnight and midday (Herdağdelen, 2013)	– be active: morning, evening and weekend (Herdağdelen, 2013)
	– be addicted to MB (Case and King, 2012)	
	– visit MB platforms more often (Pentina et al., 2014)	
	– *no difference in number of time spent on MB* (Pentina et al., 2014) – *no differences in MB access modes* (Coursaris et al., 2010)	
Use in Time	– continue using MB (Pentina et al., 2014; Wakefield et al., 2011)	– have used MB longer (Lasorsa, 2012)

Table 2. Gender differences in adoption of microblogging.

This result is also reflected in research (see Table 2): Far more studies report that females are more likely to use MB and to post there (e.g. Cheng et al., 2009). Moreover, females are also slightly more likely to be addicted to Twitter (Case and King, 2012), and plan to continue using it (Pentina et al., 2014; Wakefield et al., 2011). Together, these insights suggest that females readily embrace MB functionality as a means to maintain contact, share and discuss (Java et al., 2007).

[16] *Citations selected in italics across tables refer to Sina Weibo, otherwise Twitter.*

Nonetheless, in the case of special interest groups such as political candidates and managers, studies find no gender differences in the likelihood to adopt MB (Vergeer and Hermans, 2013; Wamba and Carter, 2013), suggesting that both groups rely on Twitter for broadcasting their information and opinions in these contexts.

4.2. Gender Differences in the Content of Microblogs

In their offline communication women are known to be more supportive and social (Cross and Madson, 1997); exhibit empathy (Clark and Bittle, 1992); concentrate on home and family (Haas, 1979), and subjective aspects (Broverman et al., 1968). In contrast, men have more pressure to establish their social standing, e.g. by communicating symbols of their success (Gefen and Ridings, 2005). As a result, male communication offline is less conducive to emotional support, with men rather exchanging facts, information and quantitative evidence (Tannen, 1994).

Evidence collected in the MB context is generally in line with these gender expectations (Table 3). We observe that personal content, lifecasting and concerns are common topics for female MB users (e.g. Lasorsa, 2012; Li et al., 2013a). This is true even in professional settings, with female journalists providing significantly more information on their day-to-day activities in their postings (Lasorsa, 2012). Furthermore, even though both groups tweet about their partners (Ikeda et al., 2013), there is evidence that females do it more often (Herdağdelen and Baroni, 2011). Moreover, women also engage more with such traditionally "female" topics as housework and food (Ikeda et al., 2013) whereas men rather emphasize a more general coverage (Holmberg and Hellsten, 2014), including such topics as politics (Cheong and Lee, 2009), serious and social topics (Li et al., 2013a), environmental news (Cheong and Lee, 2009), and events (Guan et al., 2014). Politics is particularly popular, providing evidence that male microbloggers strive to have an impact beyond their local social networks and a family unit (Cheong and Lee, 2009; Cheong et al., 2012). Nonetheless, political engagement of women is also visible: Analysis of tweets related to London riots found that even though more men (112,052) contributed to the discussion on Twitter, 80,417 women did so as well (Cheong et al., 2012). Moreover, research on Egyptian protesters reported that women were more likely to use Twitter for communicating about protests than males (Tufekci and Wilson, 2012). Nonetheless, just as it is the case of traditional blogging, female "*social power is significantly weaker than [that of] men*" (88, p. 550). Possible reasons include women's lack of interest in politics, and unwillingness of top microbloggers to re-share female posts (Harp and Tremayne, 2006).

Expected gender differences in reaction to disasters can be also observed on Twit-ter. Specifically, analysis of tweets relating to the Hurricane Irene revealed that the words "safe" and "praying" were among the top terms for women (Mandel et al., 2012), signaling their emotional needs in such situations. In contrast, men were more likely to mention "media", "breaking", "Obama" (hence reporting news and politics) and discuss practical issues such as "rooftoproofing" (Mandel et al., 2012), which is in line with their tendency to re-port and respond to calls for specific actions (Lachlan et al., 2014).

All in all, it is noteworthy that even though MB offers users significant capabilities to compensate for the gender constraints they may experience offline (e.g. by enabling more help-seeking for men, or more political and social broadcasting for women), much of this potential still remains to be utilized.

<table>
<tr><th></th><th>Females more likely to share about:</th><th>Males more likely to share about:</th></tr>
<tr><td rowspan="2">General Communication</td><td>– politics (protest) (Tufekci and Wilson, 2012)
– personal content (Lasorsa, 2012; Li et al., 2013a; Walton and Rice, 2013)
– significant others and partners (Herdağdelen and Baroni, 2011)
– housework and food (Ikeda et al., 2013)
– job (Lasorsa, 2012)
– Grey's Anatomy, Revolverheld (Cheong and Lee, 2009)
– menstruation (Thornton, 2013)
– “me now” messages (Naaman et al., 2010)
– references to gender (Burger et al., 2011)
– provide more external links (Lasorsa, 2012)
– seek help (Jackson and Lilleker, 2011)</td><td>– politics (Cheong and Lee, 2009; Cheong et al., 2012; Ikeda et al., 2013)
– serious and social topics (Li et al., 2013a)
– environmental news and issues (Cheong and Lee, 2009)
– hot social events (Guan et al., 2014; Li et al., 2013a)
– electronics (Ikeda et al., 2013)
– sports (Lasorsa, 2012)
– work (Ikeda et al., 2013)
– achievements and abilities (Jackson and Lilleker, 2011)
– sobriety checkpoints (Seitz et al., 2012)
– sales promotions (Li et al., 2013a)
– named entities (Apple's, NBA) (Bamman et al., 2014)</td></tr>
<tr><td colspan="2">– both journalist groups cover such topics as politics and government; technology and science economy and business; entertainment and celebrities; social welfare; express major opinions and disseminate information (Lasorsa, 2012)
– both mention significant other (Ikeda et al., 2013)
– help-seeping for men almost as common (Li et al., 2013a)</td></tr>
<tr><td rowspan="2">Disaster</td><td>– concern over magnitude (Lachlan et al., 2014; Mandel et al., 2012)
– dread / risk aversion (Lachlan et al., 2014)
– concern of own health (Lachlan et al., 2014)</td><td>– reference news (Mandel et al., 2012)
– reference politics (Mandel et al., 2012)
– jokes (Mandel et al., 2012)</td></tr>
<tr><td colspan="2">– no difference for likelihood to express concern for loss of material assets, concerns regarding health, or simple ambiguous fear (Lachlan et al., 2014)</td></tr>
<tr><td rowspan="2">Climate Change</td><td>– convinced users (Holmberg and Hellsten, 2014)
– specific hashtags (Holmberg and Hellsten, 2014)
– private persons (Holmberg and Hellsten, 2014)
– campaigns, movements (Holmberg and Hellsten, 2014)</td><td>– users with skeptical stance (Holmberg and Hellsten, 2014)
– general hashtags (Holmberg and Hellsten, 2014)
– climate scientists (Holmberg and Hellsten, 2014)</td></tr>
<tr><td colspan="2">– news on climate change shared by both (Holmberg and Hellsten, 2014)</td></tr>
</table>

Table 3. Gender differences in the content of microblogging.

4.3. Gender Differences in the Audience in Microblogging

Summarized in Table 4, scientific evidence on user interaction in MB suggests a complex and a somewhat convoluted picture. We observe that female sociability and concentration on close social networks gets reflected in their *reposting* behavior (Cross and Madson, 1997), with female Weibo users being more likely to repost messages (Fu and Chau, 2013), especially when in a two-way relationship, or when a message originates from friends (Luo et al., 2012). In contrast, male users are more likely to create original posts (e.g., related to hot social events) than repost information (Guan et al., 2014). When doing so, however, men rather repost novel information, and posts coming from non-friends (Luo et al., 2012). This signals their broader orientation and lesser focus on their own social network.

A somewhat different picture can be observed in terms of the *following* activity. While a study on Weibo finds that men have larger networks of followees (Zhang et al., 2013), four studies report no difference in the size of male and female networks on Twitter (Bontcheva et al., 2013; Heil and Piskorski, 2009; Lasorsa, 2012; Vergeer and Hermans, 2013). This is a note-worthy indicator for female audiences, suggesting that women are ready to embrace larger social networks (at least on Twitter) – an important affordance of MB.

	Females more likely to:	**Males more likely to:**
Reposting	– repost (*Fu and Chau, 2013*) – repost friends, superstars (*Luo et al., 2012*) – repost in 2-way relationship (*Luo et al., 2012*)	– repost novel information (*Luo et al., 2012*) – repost non-friends (*Luo et al., 2012*) – repost in 1-way relationship (*Luo et al., 2012*)
	– *no differences in retweeting frequency* (Rao et al., 2010)	
Following	– men have more followings (*Zhang et al., 2013*) – men have more reciprocated relationships (Heil and Piskorski, 2009) – men have a larger sum of followers + followees (*Wang et al., 2013*) – *no gender difference in the number of followees* (Bontcheva et al., 2013; Heil and Piskorski, 2009; Lasorsa, 2012; Vergeer and Hermans, 2013) – *no difference in follower-followee ratio* (Rao et al., 2010)	
Being reposted | commented	– get more reshares (Crigler and Just, 2013) – have a non-zero count of reposts and comments *Fu and Chau, 2013*) – get a higher response rate (*Zhang, 2012*) – have posts diffused quicker (*Li et al., 2013b*)	– get more comments (*Fu and Chau, 2013*; Xiao et al., 2012)
	– *gender has no significant effect on reposting likelihood* (*Guan et al., 2014*)	
Being followed	– men have more followers / Heil and Piskorski, 2009; *Wang et al., 2013*; *Yu and Zhu, 2012*; *Zhang et al., 2013*) – men and women more likely to follow another man (Heil and Piskorski, 2009) – *no gender difference in the number of followers* (Bontcheva et al., 2013; Lasorsa, 2012; Rao et al., 2010; Vergeer and Hermans, 2013)	
Credibility and Influence	– 70% of the amount of public event influence is contributed by males (*Li et al., 2012*) – men perceived as providing better content and be more authoritative (Pal and Counts, 2011) – male tweets perceived as more credible (esp. for politics) (*Yang et al., 2013*) – men nearly twice more influential than female users (*Zhang et al., 2013*) – two thirds of the hot Weibos are created by male users (*Li et al., 2013a*) – *no differences for behaviors of initiating / attracting communication* (Ullrich et al., 2010)	
Homo -phily	– present: reposting (*Li et al., 2013a*); interaction (Ullrich et al., 2010); commenting (Xiao et al., 2012); following (Heil and Piskorski, 2009); tie formation (for mediators) (De Choudhury, 2011) – absent: response to information seeking (*Zhang, 2012*)	

Table 4. Gender differences in interaction behavior and audience perceptions in MB.

We also observe complex dynamics in the attention paid to male and female users in terms of *following* them as opposed to *resharing / commenting* their posts. On the one hand, female politicians get more reshares on Twitter, as one study suggests (Crigler and Just, 2013). Female Weibo users also are less likely to have a zero count of reshares and comments on their posts (Fu and Chau, 2013), and their posts get diffused quicker (Li et al., 2013b). At the same time, men are shown to get more comments on both platforms (Fu and Chau, 2013; Xiao et al., 2012). Moreover, four studies show that male users have more followers on both Twitter (Heil and Piskorski, 2009) and Weibo (Wang et al., 2013; Yu and Zhu, 2012; Zhang et al., 2013), with both men and women more likely to follow other men (Heil and Piskorski, 2009).

While this evidence is questioned by studies that show no gender difference in the number of followers (Bontcheva et al., 2013; Lasorsa, 2012; Rao et al., 2010; Vergeer and Hermans, 2013), there is solid support for men enjoying greater credibility and influence on Twitter (Yang et al., 2013) and Weibo (Zhang et al., 2013), especially in a political context (Yang et al., 2013). Several reasons may underlie this: First, unattractive topic choice, with women more likely to post about personal affairs (Lasorsa, 2012; Li et al., 2013a), may dictate the narrow impact of female postings. Second, existing misbalance in perceptions can be attributed to the early advantage men hold (Gregg, 2006): Historically occupying stronger positions in society men may find it easier to establish themselves as authoritative. These forces may intervene with the perceptions and behavior of users. All in all, however, the overall picture of user interaction with regard to gender remains visibly complex and more research is needed to gain a full understanding of this phenomenon.

4.4. Gender Differences in the Motivation to Use Microblogging

When it comes to motivation to use MB two different reasons are prevalent in literature: sharing subjective information about the self and distributing objective in-formation (Pentina et al., 2014). Along these main motives are other reasons such as the interaction with others to seek help, give advice and discuss (Java et al., 2007). Further, some MB users are motivated by possibilities of professional development (Pentina et al., 2014), self-expression (Pentina et al., 2014), entertainment and leisure (Pentina et al., 2014; Johnson and Yang, 2009), emotional aspects (Zhang and Pentina, 2012; Zhao and Rosson, 2009), status enhancement (Zhan and Pentina, 2012), and educational purposes (Wakefield et al., 2011). Motives to follow a particular user have also been discussed (Clavio and Kian, 2010). Nonetheless, only few findings shed light on the moderating role of gender – an unexpected conclusion. One study suggests that gender differences in the motivation may also depend on cultural aspects (Pentina et al., 2014). Specifically, there were no gender differences in motives to use MB in the US sample, yet in the Ukrainian sample men were more likely to support their professional development via MB whereas women were more likely to use MB for entertainment, as a diary function and for expressing emotions (Pentina et al., 2014). Other studies suggest that females are more likely to ask for help (Jackson and Lilleker, 2011) and appreciate MB as a learning environment (Wakefield et al., 2011). However, research remains limited, calling for more studies in this domain.

4.5. Gender Differences in the Presentation in Microblogging

Women and men express themselves in different ways offline: The verbal language used by females is perceived to be more pleasant, polite and personal (Haas, 1979; Mulac and Lundell, 1980) whereas men express themselves in a more direct and factual fashion (Haas, 1979; Tannen, 1994). In the nonverbal domain eye contact is perceived as a friendly attitude for women but may be seen as an attempt to dominate for men (Levine and Feldman, 2002). This hints that nonverbal communication is more important for and to women and that they are likely to be more conscious in this regard (Burgoon, 1994; Levine and Feldman, 2002). Several of these particularities can be also observed in the context of MB (Table 5). Female users are more expressive in their communication and are more likely to use exclamations and question marks (e.g. Bamman et al., 2012), repetitions of characters in their preferred assessment and negation terms (e.g., Bamman et al., 2012; Bamman et al., 2014), and emoticons (e.g. Bamman et al., 2012; Bamman et al., 2014). Female users also have a more personal writing style, which is reflected in their increased use of (personal) pronouns (e.g., Bamman et al., 2012; Bamman et al., 2014), whereas men rather prefer demonstrative pronouns (Soedjono, 2012). Further, female messages are more polite and friendly as they are more likely to express a positive sentiment (Walton and Rice, 2013) and concern overall (Lachlan et al., 2014; Mandel et al., 2012).

	Females are more likely to use:	Males are more likely to use:
Abbreviations	– abbreviations (Bamman et al., 2012; Bamman et al., 2014) – OMG and LOL (Bamman et al., 2012; Bamman et al., 2014; Rao et al., 2010); haha (Burger et al., 2011)	– LMFAO (Rao et al., 2010)
	– *no difference in use of abbreviations* (Soedjono, 2012)	
Character Change	– repetitions of alphabetical characters (Bamman et al., 2012; Bamman et al., 2014; Mosquera and Moreda, 2014; Rao et al., 2010)	– alphabetical character replacements and deletions (Mosquera and Moreda, 2014)
Hashtags	– declarative (Cunha et al., 2014) and specific (Holmberg and Hellsten, 2014) hashtags	– more hashtags (Cunha et al., 2014) – imperative (Cunha et al., 2014), descriptive (Holmberg and Hellsten, 2014) hashtags
Layout	– own layout designs (Alowibdi et al., 2013b; Fortman-Roe, 2013) – magenta (Fortman-Roe, 2013), pink, yellow, green, red, light blue (Alowibdi et al., 2013a) – high brightness colors (Fortman-Roe, 2013)	– pre-defined designs (Alowibdi et al., 2013a; Alowibdi et al., 2013b) – (dark) blue (Alowibdi et al., 2013a; Fortman-Roe, 2013), black, brown, orange, gray (Alowibdi et al., 2013a) – low brightness colors (Fortman-Roe, 2013)
Linguistics	– (personal) pronouns (Bamman et al., 2012; Bamman et al., 2014: Sanders, 2012; Soedjono, 2012)	– demonstrative pronouns (Soedjono, 2012)
	– *equal use of third person pronouns* (Soedjono, 2012) – *no differences in use of articles, determiners, prepositions* (Bamman et al., 2012; Bamman et al., 2014)	
Punctuation	– exclamation and question marks (Bamman et al., 2012; Bamman et al., 2014: Rao et al., 2010)	–
Emoticons	– emoticons (Bamman et al., 2012; Bamman et al., 2014: Burger et al., 2011; Rao et al., 2010; Volkova et al., 2013) – :), <3 (Rao et al., 2010) / :D, ;) (Bamman et al., 2012; Bamman et al., 2014)	– – :D, ;) (Rao et al., 2010) / :-o, :-& (Volkova et al., 2013)
Special Words	– assessment: okay, yes[ssss]) (Bamman et al., 2012; Bamman et al., 2014) – negation (cannot, nooo[o]) (Bamman et al., 2012; Bamman et al., 2014) – non-dictionary words (Bamman et al., 2012; Bamman et al., 2014) – hesitation / backchannel sounds, e.g. ugh, grr, ah, hm (Bamman et al., 2012; Bamman et al., 2014; Rao et al., 2010)	– assessment (yessir, yea[h]) (Bamman et al., 2012; Bamman et al., 2014; Rao et al., 2010) – negation (nah, nobody, ain't) (Bamman et al., 2012; Bamman et al., 2014) – dictionary terms (Bamman et al., 2012; Bamman et al., 2014) – named entities e.g. NBA (Bamman et al., 2012; Bamman et al., 2014)
	– *men and women used nearly the same top ten words* (Sanders, 2012)	
Sentiment	– positive valence (Walton and Rice, 2013) – risk aversion (Lachlan et al., 2014) – concern (Lachlan et al., 2014; Mandel et al., 2012)	–
	– *no differences in expressing anger* (Lachlan et al., 2014) – *no difference in tone of comments* (Crigler and Just, 2013)	
Swearing	– alphabetical character change in case of swear words (Soedjono, 2012)	– swear words (Bamman et al., 2012; Bamman et al., 2014) in a homogenous writing style (Soedjono, 2012)
	– *men and women use almost the same swear words* (Soedjono, 2012)	
Tweet Style	– ellipses (Bamman et al., 2012; Bamman et al., 2014; Rao et al., 2010) – gender-marked language (Palmer, 2014)	– more full hyperlinks (Burger et al., 2011)

Table 5. Gender differences in the presentation style and layout in microblogging.

When it comes to nonverbal communication women are more likely to choose their own layout design (Alowibdi et al., 2013b; Fortman-Roe, 2013) preferring more bright and "female" colors (Alowibdi et al., 2013a; Fortman-Roe, 2013), whereas males do not put much effort into their layout design (Alowibdi et al., 2013a; Alowibdi et al., 2013b) and prefer dark and typically "male" colors (Alowibdi et al., 2013a; Fortman-Roe, 2013). Overall, both men and women behave in stereo-typical ways in MB in terms of their presentation style.

5. Concluding Remarks, Limitations and Future Research

Our literature review has identified gender differences and similarities in several aspects of MB, which are both consistent and divergent from the traditional view on gender offline (e.g., Cross and Madson, 1997; Tannen, 1994). While male microbloggers were the earliest adopters (Lasorsa, 2012), females outnumber them by now (Alexa, n.d.). Overrepresentation of female users, their desire to post more (Burger et al., 2011) and to continue using the site (Pentina et al., 2014) signals that MB both taps into their relational orientation as well as opens them new venues to reach beyond their traditional boundaries, e.g., we find that women increasingly blog in typically “male” contexts (e.g. London Riots (Cheong et al., 2012) or Egyptian protests (Tufekci and Wilson, 2012)). Nevertheless, men continue to dominate political MB-sphere. Since equal gender participation in social and political life is important for equitable and fair society, encouraging female participation in this area appears to be a critical conclusion of our research.

Based on our findings, both male and female users emerge as important population groups for marketers, yet in different ways. Since females are more likely to reshare content in their personal circles (Luo et al., 2012), they are in a strong position to create word-of-mouth through their networks, which is of interest to marketers who strive to capital-ize on the “Twitter effect” (Chu and Kim, 2011). At the same time, male opinions are perceived as more credible (Pal and Counts, 2011; Yang et al., 2013), suggesting that male endorsements are likely to have a more pronounced influence on the audience.

Our study is prone to several limitations. Due to strict space restrictions only two platforms - Twitter and Weibo – were in the focus of our attention. Yet, insights from other platforms, such as Tumblr and Yammer, may enrich the body of knowledge presented above. Further, there are some cultural and functional differences between Twitter and Weibo: For example, it is possible to express much more in 140 characters in Chinese than in English, complicating objective comparisons. Furthermore, identifying the gender of users on MB platforms might be rather challenging. Several studies use name lists to assign a gender to provided user names. This method might be prone to errors and states therefore a limitation. Finally next to the demographic dimension gender, a further analysis including the age of MB users was not part of our study. Therefore the segmentation into age groups will be part of further research. Since Twitter has a more Western and Weibo a more Eastern background, cultural differences should to be taken into account. Indeed, cultural aspects may influence the way users interact with MB platforms, calling for more studies in this area.

Since identifying areas of future research is among the main tasks of a sound meta review (Levy and Ellis, 2006; Webster and Watson, 2002) our analysis reveals a solid body of knowledge on gender differences in the areas of MB adoption, shared content, and stylistic presentation. At the same time, a rather convoluted picture of female and male interaction patterns is uncovered (“audience” category) – an important area for future investigation, especially in the context of Twitter. Moreover, we observe little or no findings on two topics of critical interest for MB providers and scholars – gender differences in motivational patterns and privacy behavior of MB users. Together, these under-researched domains offer exciting opportunities for future scholarly endeavors.

References from Review

Alowibdi, J.S., Buy, U.A., Yu, P.: Empirical Evaluation of Profile Characteristics for Gen-der Classification on Twitter. In: 12th International Conference on Machine Learning and Applications. IEEE, vol. 1, pp. 365-369 (2013a)

Alowibdi, J.S., Buy, U.A., Yu. P.: Language Independent Gender Classification on Twit-ter. In: International Conference on Advances in Social Networks Analysis and Mining. IEEE/ACM, pp. 739-743 (2013b)

Bamman, D., Eisenstein, J., Schnoebelen, T.: Gender in Twitter: Styles, Stances, and So-cial Networks. arXiv:1210.4567 (2012)

Bamman, D., Eisenstein, J., Schnoebelen, T.: Gender Identity and Lexical Variation in So-cial Media. Journal of Sociolinguistics, 18(2), 135-160 (2014)

Bontcheva, K., Gorrell, G., Wessels, B.: Social Media and Information Overload: Survey Results. arXiv:1306.0813 (2013)

Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating Gender on Twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 1301-1309 (2011)

Case, C.J., King, D.L.: Exploring the Undergraduate World of Electronic Social Network-ing: An Examination of Attitude, Adoption, and Usage. Issues in Information Systems 13(1), 51-58 (2012)

Cheng, A., Evans, M., Singh, H.: Inside Twitter: An In-depth Look Inside the Twitter World. Report of Sysomos, June, Toronto, Canada (2009)

Cheong, M., Lee, V.: Integrating Web-based Intelligence Retrieval and Decision-making from the Twitter Trends Knowledge Base. In: Proceedings of the 2nd ACM Workshop on Social Web Search and Mining. ACM, pp. 1-8 (2009)

Cheong, M., Ray, S., Green, D.: Interpreting the 2011 London Riots from Twitter Metada-ta. In: Proceedings of the 12th International Conference on Intelligent Systems Design and Applications, pp. 915-920 (2012)

Clavio, G., Kian, T.M.: Uses and Gratifications of a Retired Female Athlete’s Twitter Fol-lowers. International Journal of Sport Communication, 3(4), 485-500 (2010)

Coursaris, C.K., Yun, Y., Sung, J.: Understanding Twitter’s Adoption and Use Continu-ance: The Synergy Between Uses and Gratifications and Diffusion of Innovations. In: SIGHCI 2010 Proceedings (2010)

Crigler, A.N., Just, M.R.: Gender and Self-Presentation in Social Media: An Analysis of Four Competitive 2012 US Senate Races. In: APSA 2013 Annual Meeting Paper (2013)

Cunha, E., Magno, G., Gonçalves, M.A., Cambraia, C., Almeida, V.: He Votes or She Votes? Female and Male Discursive Strategies in Twitter Political Hashtags. PloS one, 9(1) (2014)

Dar, H., Shah, A.: Analysis of SNS Popularity from Different Perspectives Among Users. In: 5th International Conference on Information and Communication Technology for the Muslim World (ICT4M). IEEE, pp. 1-4 (2013)

De Choudhury, M.: Tie Formation on Twitter: Homophily and Structure of Egocentric Networks. In: Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and Third International Conference on Social Computing (SocialCom). IEEE, pp. 465-470 (2011)

Fortmann-Roe, S.: Effects of Hue, Saturation, and Brightness on Color Preference in So-cial Networks: Gender-based Color Preference on the Social Networking Site Twitter. Color Research & Application 38(3), 196-202 (2013)

Fox, S., Zickuhr, K., Smith, A.: Twitter and Status Updating Fall 2009. Pew Internet & American Life Project. http://www.pewinternet.org/Reports/2009/17-Twitter-and-Status-Updating-Fall-2009.aspx (2009)

Fu, K., Chau, M.: Reality Check for the Chinese Microblog Space: A Random Sampling Approach. PloS one 8(3) (2013)

Guan, W., Gao, H., Yang, M., Yuan, L., Ma, H., Qian, W., Cao, Z., Yang, X.: Analyzing User Behavior of the Micro-blogging Website Sina Weibo During Hot Social Events. Physica A: Statistical Mechanics and its Applications 395, 340-351 (2014)

Hargittai, E., Litt, E.: The Tweet Smell of Celebrity Success: Explaining Variation in Twitter Adoption Among a Diverse Group of Young Adults. New Media & Society 13(5), 824-842 (2011)

Heil, B., Piskorski, M.: New Twitter Research: Men Follow Men and Nobody Tweets. Harvard Business Review 1 (2009)

Herdağdelen, A.: Twitter n-gram Corpus with Demographic Metadata. Language Re-sources and Evaluation 47(4), 1127-1147 (2013)

Herdağdelen, A., Baroni, M.: Stereotypical Gender Actions can be Extracted from Web Text. J of the Am Soc for Info Sci and Tech 62(9), 1741-1749 (2011)

Holmberg, K., Hellsten, I.: Analyzing the Climate Change Debate on Twitter: Content and Differences Between Genders. In: Proceedings of the 2014 ACM Conference on Web Sci-ence, pp. 287-288. ACM (2014)

Ikeda, K., Hattori, G., Ono, C, Asoh, H., Hihashino, T.: Twitter User Profiling Based on Text and Community Mining for Market Analysis. Knowledge-Based Systems 51, 35-47 (2013)
Jackson, N., Lilleker, D.: Microblogging, Constituency Service and Impression Manage-ment: UK MPs and the Use of Twitter. The Journal of Legislative Studies 17(1), 86-105 (2011)
Lachlan, K.A., Spence, P.R., Lin, X.: Expressions of Risk Awareness and Concern Through Twitter: On the Utility of Using the Medium as an Indication of Audience Needs. Computers in Human Behavior 35, 554-559 (2014)
Lasorsa, D.: Transparency and Other Journalistic Norms on Twitter: The Role of Gender. Journalism Studies 13(3), 402-417 (2012)
Lenhart, A., Purcell, K., Smith, A., Zickuhr, K.: Social Media & Mobile Internet Use Among Teens and Young Adults. Pew Internet & American Life Project (2010)
Li, X., Xie, Y., Li, C., Sun, X.: Analyzing the Public Events' Influence Via Open Mi-croblogging APIs. In: International Conference on Machine Learning and Cybernetics (ICMLC). IEEE, vol. 1, pp. 84-90 (2012)
Li, Y., Gao, H., Yang, M., Guan, W., Ma, H., Qian, W., Cao, Z., Yang, X.: What are Chi-nese Talking About in Hot Weibos? arXiv:1304.4682 (2013a)
Li, Z., Lv, T., Zhang, X., Chen, X.: The Effects of Personal Characteristics and Interper-sonal Influence on Privacy Information Diffusion in SNS. In: International Conference on Service Operations and Logistics, and Informatics (SOLI). IEEE, pp. 413-418 (2013b)
Luo, Z., Wu, X., Cai, W., Peng, D.: Examining Multi-factor Interactions in Microblogging Based on Log-linear Modeling. In: Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012). IEEE Computer Society, pp. 189-193 (2012)
Mandel, B., Culotta, A., Boulahanis, J., Stark, D., Lewis, B., Rodrigue, J.: A Demographic Analysis of Online Sentiment During Hurricane Irene. In: Proceedings of the Second Workshop on Language in Social Media. Association for Computational Linguistics, pp. 27-36 (2012)
Mislove, A., Lehmann, S., Ahn, Y.Y., Onnela, J.P., Rosenquist, J.N.: Understanding the Demographics of Twitter Users. In: 5th International Conference on Weblogs and Social Media (2011)
Mosquera, A., Moreda, P.: Mining Lexical Variants from Microblogs: An Unsupervised Multilingual Approach. In: Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM). EACL, pp. 1-7 (2014)
Naaman, M., Boase, J., Lai, C.-H.: Is it Really About Me?: Message Content in Social Awareness Streams. In: Proceedings of the Conference on Computer Supported Coopera-tive Work. ACM, pp. 189–192, New York, USA (2010)
Pal, A., Counts, S.: What's in a@ Name? How Name Value Biases Judgment of Microblog Authors. In: International Conference on Weblogs and Social Media (2011)
Palmer, J.: The Role of Gender on Social Network Websites. Stylus Knights Write Show-case Special Issue (2014)
Pentina, I., Basmanova, O., Zhang, L.: A Cross-national Study of Twitter Users' Motiva-tions and Continuance Intentions. Journal of Marketing Communications, 1-20 (2014)
Rao, D., Yarowsky, D., Shreevats, A., Gupta, M.: Classifying Latent User Attributes in Twitter. In: Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents. ACM, pp. 37-44 (2010)
Sanders, E.P.: Collecting and Analysing Chats and Tweets in SoNaR. In: European Lan-guage Resources Association (ELRA), Istanbul, Turkey (2012)
Seitz, C. M., Orsini, M. M., Fearnow-Kenney, M., Hatzudis, K., Wyrick, D. L.: Twitter as a Tool to Warn Others About Sobriety Checkpoints: A Pilot Observational Study. Interna-tional Electronic Journal of Health Education 15, 112-119 (2012)
Soedjono, A.H.: The Comparisons Between the Language Used by Male and Female Peers in Twitter. Anglicist 1, 1-6 (2012)
Thornton, L.J.: "Time of the Month" on Twitter: Taboo, Stereotype and Bonding in a No-Holds-Barred Public Arena. Sex Roles 68(1-2), 41-54 (2013)
Tufekci, Z., Wilson, C.: Social Media and the Decision to Participate in Political Protest: Observations from Tahrir Square. Journal of Communication 62(2), 363-379 (2012)
Ullrich, C., Borau, K., Stepanyan, K.: Who Students Interact With? A Social Network Analysis Perspective on the Use of Twitter in Language Learning. In: Sustaining TEL: From Innovation to Learning and Practice, pp. 432-437. Springer, Heidelberg (2010)
Vergeer, M. Hermans, L.: Campaigning on Twitter: Microblogging and Online Social Networking as Campaign Tools in the 2010 General Elections in the Netherlands. Journal of Computer-Mediated Communication 18, 399–419 (2013)
Volkova, S., Wilson, T., Yarowsky, D.: Exploring Demographic Language Variations to Improve Multilingual Sentiment Analysis in Social Media. In: EMNLP, pp. 1815-1827 (2013)

Wakefield, J.S., Warren, S.J., Alsobrook, M.: Learning and Teaching as Communicative Actions: A Mixed-methods Twitter Study. Knowledge Management & E-Learning: An In-ternational Journal (KM&EL) 3(4), 563-584 (2011)
Walton, C.S., Rice, R.E.: Mediated Disclosure on Twitter: The Roles of Gender and Iden-tity in Boundary Impermeability, Valence, Disclosure, and Stage. Computers in Human Behavior 29(4), 1465-1474 (2013)
Wamba, F.S., Carter, L.: Twitter Adoption and Use by SMEs: An Empirical Study. In: The 46th Hawaii International Conferences on System Sciences (HICSS), Maui, Hawaii (2013)
Wang, L., Qu, W., Sun, X.: An Analysis of Microblogging Behavior on Sina Weibo: Per-sonality, Network Size and Demographics. In: Cross-Cultural Design. Methods, Practice, and Case Studies, pp. 486-492. Springer Berlin Heidelberg (2013)
Xiao, C., Su, L., Bi, J., Xue, Y., Kuzmanovic, A.: Selective Behavior in Online Social Networks. In: International Conferences on Web Intelligence and Intelligent Agent Tech-nology (WI-IAT). IEEE/WIC/ACM, vol. 1, pp. 206-213 (2012)
Yang, J., Counts, S., Morris, M.R., Hoff, A.: Microblog Credibility Perceptions: Compar-ing the USA and China. In: Proceedings of the 2013 Conference on Computer Supported Cooperative Work. ACM, pp. 575-586 (2013)
Yu, Y., Zhu, Y.: Follow Me: An Analysis of Self-traits, Motivation, Microblog Usage and Attractiveness of One's Microblog. In: 9th International Conference on Service Systems and Service Management (ICSSSM). IEEE, pp. 148-153 (2012)
Zhang, P.: Information Seeking Through Microblog Questions: The Impact of Social Capi-tal and Relationships. In: Proceedings of the American Society for Information Science and Technology 49(1), pp. 1-9 (2012)
Zhang, X., Gao, Q., Khoo, C.S.G., Wu, A.: Categories of Friends on Social Networking Sites: An Exploratory Study. In: The 5th International Conference on Asia-Pacific Library and Information Education and Practice (A-LIEP 2013), pp. 244-259. Khon Kaen City, Thailand (2013)

Other References

Alexa.com, http://www.alexa.com/ (Accessed: 02.08.2014).
Broverman, D.M., Klaiber, E.L., Kohayashi, Y., Vogel, W.: Roles of Activation and Inhi-bition in Sex Differences in Cognitive Abilities. Psychological Review 75(1), pp. 23-50 (1968)
Burgoon, J.K.: Nonverbal signals. in Handbook of Interpersonal Communication, M. L. Knapp and G.R. Miller (eds.) , 2nd ed. Thousand Oaks (CA): Sage, pp. 229–85 (1994)
Chen, S., Zhang, H., Lin, M., Lv, S.: Comparision of microblogging service between Sina Weibo and Twitter. In: International Conference on Computer Science and Network Tech-nology (ICCSNT). IEEE, vol. 4, pp. 2259-2263 (2011)
Chu, S.C., Kim, Y.: Determinants of Consumer Engagement in Electronic Word-of-Mouth (eWOM) in Social Networking Sites. International Journal of Advertising 30(1), 47-75 (2011)
Clark, M.L., Bittle, M.L: Friendship Expectations and the Evaluation of Present Friend-ships in Middle Childhood and Early Adolescence. Child Study Journal (1992)
Cross, S. E., Madson, L.: Models of the Self: Self-Construals and Gender. Psychological Bulletin 122(1), 5-37 (1997)
Eagly, A.H.: Sex Differences in Social Behavior: A Social-Role Interpretation. Lawrence Erlbaum Associates Inc. New York Hillsdale (1987)
Gefen, D., Ridings, C.M.: If You Spoke as She Does, Sir, Instead of the Way You Do: A Sociolinguistics Perspective of Gender Differences in Virtual Communities. SIGMIS Da-tabase 36(2), 78-92 (2005)
Gregg, M.C.: Posting with Passion: Blogs and the Politics of Gender. (2006)
Haas, A.: Male and Female Spoken Language Differences: Stereotypes and Evidence. Py-chological Bulletin 86(3), 616-626 (1979)
Hafkin, N.J.: Women and Gender in ICT Statistics and Indicators for Development. The MIT Press 4(2), 25-41 (2008)
Harp, D., Tremayne, M.: The Gendered Blogosphere: Examining Inequality Using Net-work and Feminist Theory. Journalism & Mass Communication Quarterly 83(2), 247-264 (2006)
Huffaker, D.A., Calvert, S.L.: Gender, Identity, and Language Use in Teenage Blogs. Journal of Computer-Mediated Communication 10(2) (2005)
Humphreys, L., Gill, P., Krishnamurthy, B.: How Much is Too Much? Privacy Issues on Twitter. In: Conference of International Communication Association, Singapore (2010)
Jackson, L.A., Ervin, K.S., Gardner, P.D., Schmitt, N.: Gender and the Internet: Women Communicating and Men Searching. Sex Roles 44(5-6), 363-379 (2001)
Java A, Song X, Finin T, Tseng, B.: Why We Twitter: Understanding Microblogging Us-age and Communities. In: Proceedings of SNA-KDD Workshop, San Jose, CA, (2007)

Johnson, P.R., Yang, S.-U.: Uses and Gratifications of Twitter: An Examination of User Motives and Satisfaction of Twitter Use. In: Proceedings of Annual Association for Educa-tion in Journalism and Mass Communication Conference. Boston, MA, (2009)
Landis, J.R., Koch, G.G.: The Measurement of Observer Agreement for Categorical Data. Biometrics 33(1), 159-174 (1977)
Levine, S.P., Feldman, R.S.: Women and Men’s Nonverbal Behavior and Self-Monitoring in a Job Interview Setting. Applied HRM Research 7(1), 1–14 (2002)
Levy, Y., Ellis, T.J.: A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research. Informing Science 9 (2006)
Mulac, A., Lundell, T.L.: Differences in Perceptions Created by Syntactic-Semantic Pro-ductions of Male and Female Speakers. Communications Monographs 47(2), 111-118 (1980)
Statista.com, http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users (Accessed: 02.08.2014)
Stewart-Williams, S., Thomas, A.G.: The Ape That Thought It Was a Peacock: Does Evo-lutionary Psychology Exaggerate Human Sex Differences? Psychological Inquiry 24(3), 137-168 (2013)
Tannen, D.: Gender and Discourse. New York, Oxford University Press, pp. 119-121 (1994)
Venkatesh, V., Morris, M.G.: Why Don't Men Ever Stop to Ask for Directions? Gender, Social Influence, and Their Role in Technology Acceptance and Usage Behavior. MIS Quarterly 24(1), 115-139 (2000)
Webster, J., Watson, R.T.: Analyzing the Past to Prepare for the Future: Writing a Litera-ture Review. MIS Quarterly 26(2), 13-23 (2002)
Zhang, L., Pentina, I.: Motivations and Usage Patterns of Weibo. Cyberpsychology, Be-havior, and Social Networking 15(6), 312-317 (2012)
Zhao, D., Rosson, M. B.: How and Why People Twitter: The Role that Micro-blogging Plays. In: Informal Communication at Work (2009)

ARTICLE 13:

GENDER DIFFERENCES IN ONLINE GAMING: A LITERATURE REVIEW

Authors: Natasha Veltri, University of Tampa, Tampa, Florida, USA

Hanna Krasnova, Universität Potsdam, Potsdam, Germany

Annika Baumann, Humboldt-Universität zu Berlin, Berlin, Germany

Neena Kalayamthanam, Universität Bern, Bern, Switzerland

Proceedings: 20th Americas Conference on Information Systems (AMCIS 2014)

Publication year: 2014

Abstract

Online gaming continues to grow in popularity enabled by advances in web and mobile technology. Many players around the world enjoy the competitive atmosphere, mental challenges, social interaction and fantasy aspects of the games. However, gaming continues to be perceived as an activity for adolescents and males, which presents problems for companies trying to leverage games for training and marketing because it excludes half of the population – females. To better understand online gaming behavior of men and women we systematically reviewed extant literature on gaming and documented gender differences and similarities in six aspects of gaming: (1) Adoption, (2) Motivation, (3) Social Interaction, (4) Self-presentation, (5) Skills and Performance, and (6) Play. We find that there are both similarities and differences in gaming choices, motives, play behavior and performance of men and women. Based on our findings, we suggest possible strategies for developers, marketers and educators to achieve gender parity.

***Keywords**: Online Games, Virtual Worlds, MMORPG, Gender Differences, Men, Women*

1. Introduction

Played by over a billion people globally, digital games are challenging, engaging and fun (Liu et al., 2013). Online games, the most popular of digital games, are enabled by modern information and communication technologies and played on an Internet-based platform. These games transport users to a virtual environment and are played individually or together with other online gamers. Many players crave the competitive nature of online games (Liu et al., 2013), others seek hedonic gratification (enjoyment, fantasy and escape from reality), yet others relish social interaction (Li et al., 2013). The growing popularity of gaming is further amplified by the spread of mobile technologies with 36% of gamers using their mobile devices to access their favorite games anytime and anywhere in 2013 (Entertainment Software Association, 2013).

There are many types of online games. Among them, puzzle, logic and card games are most popular accounting for 34% of all online games, followed by action, sports, strategy and role-playing games (26%), casual and social games (19%) and persistent multiplayer universe games (14%) (Entertainment Software Association, 2013). An important category, Massively Multi-player Online Role-Playing Games (MMORPGs), allow players to create a new identity, navigate their avatar in a 3D environment and interact with others in a “reality-like” setting. The environment, the roles and stories to be created (Kuo at al., 2012), and the experience in a place different from the one they are physically in (Schroeder, 2006) entice gamers to return.

Enormous popularity of online gaming not only created a lucrative market for game developers, but also challenged businesses, non-profit organizations, educational institutions, political parties and governments to harness the power of virtual environments and their users. Games are increasingly used for learning and training, interacting with customers and political supporters as well as promoting products and services. However, gaming is traditionally perceived as an activity for adolescents and males (Becerra et al., 2008; Chen, 2010), which presents problems for companies trying to leverage games for training or marketing as it excludes half of the population – females. As the number of female gamers continues to grow, it is important to understand the motives and behavior of female gamers. To achieve this, more insight is needed into gender differences in behavioral patterns and perceptions of gamers.

The purpose of this study is to systematically document reported gender differences in various aspects of gaming. We believe that identifying these differences is important because gender is a fundamental characteristic that underlies the behavior and societal roles of men and women. Indeed, significant differences exist in the patterns of IT use by men and women (Venkatesh and Morris, 2000), and these differences are likely to transfer to online gaming environment. Moreover, gender represents the most simple and, at the same time, effective variable used for targeting, presenting vast implications for marketing and learning. Knowledge gained in this research will inform other scholars of the state-of-the-art in this area. Finally, our findings are likely to be valuable for game developers, educational specialists and advertisers.

2. Research Method

In conducting our literature review we followed recommendations of Webster and Watson (2002). First, we searched public scholarly databases Google Scholar, ScienceDirect, EBSCOhost, SpringerLink and JSTOR to identify relevant articles. We were interested in previously reported insights regarding gender differences in online gaming, and thus we entered combinations of words {Virtual Worlds, Online Games, MMORPG} and {Gender; Men; Women; Male; Female}. Additionally, we searched for the most popular virtual game platforms using keywords {Second Life, Habbo, World of Warcraft}. The gender search was extended to include {Girl} and {Boy} to ensure that articles examining gaming behavior of children or teenagers were in our sample. In the second step, we conducted a backward search by reviewing the citations in the identified articles to find additional studies we might have missed. Finally, a forward search was also conducted. The final sample included 47 articles published between 2004 and 2014 that reported or discussed gender differences in online gaming[17]. These articles were comprehensively reviewed to structure gender-relevant insights.

2.1. Aspects of Gaming Behavior

After careful review of the articles, we identified six areas where significant gender differences in gaming context have been reported: (1) *adoption*, (2) *motivation*, (3) *social interaction*, (4) *self-presentation*, (5) *skills and performance,* and (6) *play*. Figure 1 summarizes these themes.

Figure 1. Aspects of gender differences in online gaming.

3. Findings: Gender Differences in Online Gaming

3.1. Adoption

Traditionally, video games have belonged in the male domain (Lucas and Sherry, 2004; Fox and Tang, 2013). Existing findings continue to support this perception, albeit some exceptions (Table 1). First, men and women adoption rates differ. Indisputably, studies indicate that more men play online games than women (Becerra et al., 2008; Hainey et al., 2011). Second, men develop interest in computer games earlier in their life, while females take up gaming later in life and on average female players are older than male (e.g. Hainey et al., 2011).

[17] A complete list of articles is available from the authors upon request.

Source	Study Findings	Conclusion
Number of users		
Becerra et al. (2008)	Males use virtual worlds more than females.	More males play online games
Hainey et al. (2011)	More male players (92.6% of men vs. 69.9% of women).	
Time of Adoption		
Hainey et al. (2011)	Males have played games for a longer time period (13.68 years) than females (11.05 years).	Males adopt online games earlier in life
	Males exhibit interest in computer games earlier.	
Level of Activity (Time, Frequency)		
Chou and Tsai (2007)	Males spend more time playing (284 min/week vs. 172 min/week for females).	Males play more frequently and spend more time playing
Hainey et al. (2011)	Males spend more time playing (9.02 hours/week for males vs. 4.39 hours/week for females).	
Chen (2010)	Males spend more time gaming online.	
Cohen (2009)	Females less likely to play time-intensive games.	
Kuo et al. (2012)	Males spend less time playing (173.96 min/day for males vs. 223.70 min/day for females).	Females spend more time playing
Williams et al. (2008)	Males spend less time playing (25.03 hours/week vs. 29.31 hours/week for females).	

Table 1. Gender-relevant findings on gaming adoption.

Men and women also differ on time they spend playing and in frequency of play. Multiple studies report that men play more often and for a significantly longer time than women (Chou and Tsai, 2007; Hainey et al., 2011; Chen, 2010). However, there is no consistency in existing findings regarding game duration and several studies suggest that women spend more time gaming (Williams et al., 2008; Kuo et al., 2012). It is important to note that the sample in the study by Kuo et al. (2012) had over 3,000 players ranging in age from 8 to 86 (average 25), possibly indicating a difference in the gaming time for older players. While we can't unequivocally conclude that men spend more time playing because of these conflicting results, earlier research on game duration in offline context suggests that men play for longer periods. For example, in childhood play 72% of all boys' activities last longer than one hour, compared to only 43% of girls' activities (Lever, 1976). This indicates that boys have a tendency to play for longer periods of time, corroborating the findings of longer online game duration for men.

3.2. Motivation

Many studies investigated what motivates users to play online games. Lucas and Sherry (2004) identify six motives for gaming: competition, challenge, arousal, fantasy, diversion, and social interaction. We follow this classification to group identified gender-relevant insights in this area (Table 2). It appears that men and women exhibit some differences and similarities in the motives to play online games (Chou and Tsai, 2007; Lucas and Sherry, 2004), with research evidence offering some conflicting conclusions, as shown in Table 2.

Source	**Study Findings**	**Conclusion**
Competition		
Yee (2006)	Male players score higher on all achievement dimensions.	Achievement more important for men
Hassouneh and Brengman (2014)	Achievement factor more motivational for male users.	
Challenge		
Lucas and Sherry (2004)	Challenge is the top-rated motive for both.	Both genders seek challenge, men more so
	Young men more motivated by challenge.	
Arousal		
Choi et al. (2012)	Avatar manipulation and shopping are viewed as significant entertainment activities in virtual worlds, with female users reporting higher entertainment scores.	Both men and women recognize benefits of arousal
Zhou et al. (2011)	Females emphasize exploring, researching and shopping within the Second Life world more.	
Chou and Tsai (2007)	Entertainment, excitement and enjoyment sharing are greater motives for men to play computer games.	
Fantasy		
Chou and Tsai (2007)	Fantasy is a greater motive for men to play computer games.	Fantasy more important for men
Diversion		
Chou and Tsai (2007)	No gender difference on escapism and filling time as motives to play computer games.	No differences for escapism and pass time
Hassouneh and Brengman (2014)	Females more motivated by escapism.	Female stress escapism more
Friendship and Social Acceptance		
Yee (2006)	Women score higher on the relationship dimension (e.g. personal self-disclosure, finding and giving support).	Both men and women recognize relational benefits
Chou and Tsai (2007)	Men emphasize computer games as a social device.	
Hassouneh and Brengman (2014)	Male users seek 'Relationships' more; Females seek 'Friendship' more.	
Iqbal et al. (2010)	Need to socialize is an important reasons for girls to not participate in Virtual Worlds.	Differences in motives to reject virtual worlds
Other Motives		
Choi et al. (2012)	Females show greater interest in information seeking and derive more benefits from information resources in virtual worlds.	Females seek information; functional and experiential value
Zhou et al. (2011)	Women pay greater attention to functional and experiential values.	
	Males emphasize making money on Second Life.	Men emphasize money-making, possibility to "manipulate", seek information and wait for "good" games
Chou and Tsai (2007)	Information seeing is a greater motive for men to play computer games.	
Hassouneh and Brengman (2014)	Men motivated by "manipulation" motive.	
Iqbal et al. (2010)	Non availability of better games is an important reason for boys to not participate in Virtual Worlds.	

Table 2. Gender-relevant findings on gaming motivation.

3.3. Social Interaction

The ability to play, interact and compete with others is a major attraction for all gamers, yet there are similarities and differences in the types of interactions men and women seek (Table 3). Various studies report female gamers as more sociable: they are more likely to meet people while playing (Guadagno et al., 2011), engage in group activities in virtual environments (Choi et al., 2012), and participate more in peer-discussions (Hou, 2012). Findings also suggest that female players are more cooperative and use more supportive language to encourage their counterparts (Hong and Hwang, 2012). Further, women are also more likely to recruit new players both online and offline (Williams et al., 2006). Men appear to enjoy the social aspect of gaming as well, but they look for different things in such interactions and relationships (Yee, 2006). According to Hong and Hwang (2012) boys are less willing to seek help than girls in the beginning of the game, but do so later on as the competitive pressure mounts. A similar pattern can be also found in children offline games, where there is a much higher competitive spirit in boys' activities (Lever, 1976).

Source	**Study Findings**	**Conclusion**
Relationship Seeking		
Guadagno et al. (2011)	Women are more likely to meet people.	Women are more active in seeking friendships in online games
Hassouneh and Brengman (2014)	'Friendship seekers' are more common among females.	
Choi et al. (2012)	Female Second Life players more active in making friends (social life factor).	
Type of Relationship Seeking		
Yee (2006, p. 774)	"*Male players socialize just as much as female players, but are looking for very different things in those relationships*".	Males seek different types of relationships
Group / Social Behavior		
Hou (2012)	Female gamers more likely to engage in peer discussions.	Women are more social and participate in group activities, discussions and generally seek help more
Choi et al. (2012)	Female players of Second Life more active in group activities (social life factor).	
Hong and Hwang (2012)	Girls exhibit more communal attitude; use more encouraging language (e.g. "Go" or "Hurry up").	
	Boys more likely to use negative expressions (e.g. "You are stupid"); less willing to ask for help in the beginning of the game.	
Yee et al. (2007)	Male avatars in male-male dyads less likely to look at each other than those in female-female and mixed dyads.	
Recruitment of New Players		
Williams et al. (2006)	Female players more likely to recruit new members from online or offline.	Women recruit new players more

Table 3. Gender-relevant findings on social interaction in gaming.

3.4. Self-Presentation

The interaction in virtual online game-playing environments often takes place using avatars, which can be customized by players. Females pay more attention to the visual appearance of their avatars by regularly buying items such as clothes for them to change their look (Hou, 2012). Furthermore female players with high neuroticism scores and introverts of both genders create more attractive avatars (Dunn and Guadagno, 2012). This focus on avatar appearance is promoted by girl specific games online and offline, such as dressing-up a Barbie in real life or on Barbie.com (Iqbal et al., 2010).

While both genders are curious to try an avatar of the opposite gender, switching gender is more popular among male players (Ducheneaut et al., 2009; Hassouneh and Brengman, 2014). This gender switching behavior for males is somewhat surprising because masculine behavior is typically rewarded in gaming world (Fox and Tang, 2013) and female voiced avatars are reported to receive three times as many negative comments (Kuznekoff and Rose, 2013). A possible explanation for this behavior might be that female avatars get more attention in form of messages and friend requests (Kuznekoff and Rose, 2013) as well as an improvement in helping behavior in both directions, i.e. males are more likely to ask for help and at the same time receive more assistance (Lehdonvirta et al., 2012). Table 4 summarizes our findings.

Source	Study Findings	Conclusion
Gender Switching		
Ducheneaut et al. (2009)	24% players used the opposite gender: more male players used a female character.	Men are more likely to use avatars of the opposite sex
Hassouneh and Brengman (2014)	More males are interested in using an avatar of the opposite gender than females (10% males vs. 4% females).	
Avatar Customizing and Visual Appearance		
Hou (2012)	Women pay more attention to their avatar's appearance.	Women are more concerned about the visual appearance of their avatars
Guadagno et al. (2011)	Women are more likely to change their avatar's appearance.	
	Women are more likely to buy clothes/objects for their avatars.	
Dunn and Guadagno (2012)	Women with high neuroticism and introverts of both genders are more likely to create attractive avatars.	

Table 4. Gender-relevant findings on self-presentation.

3.5. Skills and Performance

Playing games requires different types of skills, such as spatial, cognitive, motor and social. Overall, results indicate that male players perform better in online games (at least in the types that have been studied, see Table 5). Male navigation performance is superior (Tlaukaa et al., 2005), while females demonstrate more navigational difficulties, take longer to travel from start to end and make more incorrect navigational decisions (Tippett et al., 2009). Male adolescents seem to have a longer attention span for games in general (Cohen, 2009) Furthermore the predominantly male orientation of many popular online games might support this factor. Since females are motivated by ease of use in their use of technology (Venkatesh and Morris, 2000) and prefer to play games where the skills can be acquired much faster (Lever, 1976), they may not develop the same gaming skills that men do.

Source	Study Findings	Conclusion
Gaming Skills		
Tippett et al. (2009)	Men navigate more efficiently through virtual environments.	Men exhibit better spatial skills
	Men had greater spatial problem solving efficiency than women.	
Gaming Performance		
Tlaukaa et al. (2005)	Women needed more time to travel from start to finish of the route.	Men perform better on a variety of tasks
	Women needed more time for directional estimates.	
	Women placed the target objects more imprecise on the map.	
	Women made more incorrect navigational decisions.	
	Women performed less accurately when asked to navigate back to the start location.	

Table 5. Gender-Relevant Findings on Skills and Performance

3.6. Play

Both differences and similarities exist in game preferences and play patterns of men and women (Table 6). Men play more action and simulation games, while women play logic and skills training games or do not play at all (Quaiser-Pohl et al., 2005). Female gamers usually play very gender specific games, such as dress up on Barbie.com (Iqbal et al., 2010), or family oriented simulations related to pregnancy and maternity (Lomanowska and Guitton 2013). In general, women prefer to play games where they can correct mistakes easily and undo functions (Wang, 2013).

There are also gender differences in various roles user play. Females are more 'role players', 'friendship seekers' and 'achievement seekers', whereas males are 'manipulators', 'uninvolved users' and ‘relationship seekers’ (Hassouneh and Brengman, 2013). Males are also found to seek social dominance and power over women in games (Fox and Tang, 2013).

Some similarities exist in gaming as well. When it comes to making money in virtual worlds, studies report that both women (Hassouneh and Brengman, 2014) and men enjoy it (Zhou et al., 2011). Both men Guadagno et al. (2011) and women (Hassouneh and Brengman, 2014) like to build objects and work on their virtual property. Further, both genders play the role of a ‘relationship seeker’ equally (Hassouneh and Brengman, 2014). Finally, the number of males and females found in sex-related simulations are almost the same (Lomanowska and Guitton, 2013).

Source	**Study Findings**	**Conclusion**
Types of Games		
Quaiser-Pohl et al. (2005)	More males among “action-and-simulation” game players (81.7%).	Men prefer action-oriented games
Zhang et al. (2010)	Male avatars perform more active physical actions than female avatars.	
Quaiser-Pohl et al. (2005)	More females among “logic-and-skill-training” game players (82.9%) and among non-players (81.9%).	Women play traditional female-oriented games, logic games and games with undo functionality
Iqbal et al. (2010)	Girls play gender specific games, such as dress-up.	
Lomanowska and Guitton (2013)	Female avatars were more commonly in family oriented simulations related to pregnancy and maternity.	
Wang (2013)	Ability to undo operations and easily recover from mistakes has more influence on the intent to play a new game for female online game players than for male players.	
Types of Activities in Games		
Zhou et al. (2011)	Male users take more notice of using Second Life for making money.	Men enjoy building, making money
Guadagno et al. (2011)	Men build more things, and own and work on their own virtual property.	
Zhou et al. (2011)	Females are more likely to shop and research.	Women enjoy socializing, shopping, exploring and improving their character
Guadagno et al. (2011)	Females meet people, shop, modify their avatar’s appearance and buy clothes/objects for their avatars.	
Hou (2012)	Females are more likely to configure various items and tools.	
Hassouneh and Brengman (2014)	Men and women show interest in relationship seeking.	Players of both genders seek relationships and engage in sex-related activities
Lomanowska and Guitton (2014)	Both males and females are found in sex-related simulations.	
Types of Roles		
Hassouneh and Brengman (2014)	Male users of social virtual worlds are ‘role players’ (18.6% of male users), ‘manipulators’ (17.7%), ‘uninvolved’ users (16.8), or ‘relationship seekers’ (13.2%).	Men and women take on somewhat different roles.
	Female users of social virtual worlds are ‘role players’ (21.3%), ‘friendship seekers’ (16.2%), ‘achievement seekers’ (15.3%), or ‘relationship seekers’ (7.9%).	
Negative Aspects of Gaming		
Huanhuan and Su (2013)	Males are at a greater risk of developing online game addiction than females.	Males more likely to be addicted to gaming

Table 6. Gender-relevant findings on play.

4. Discussion

Our literature review uncovered various gender differences and similarities in online gaming. First, gaming continues to be dominated by male players, who are more motivated to play, start playing games earlier in life, play more frequently and spend more time playing. Persistent perceptions of gaming as a male domain and sexist and even hostile environment in some worlds and games could be reasons for this trend (Fox and Tang, 2013). However, female participation in games continues to grow, and recent statistics indicate that women were 46% of the most frequent purchasers of video games in 2013 (Entertainment Software Association, 2013). Thus, women may simply be playing different types of games.

Second, we find that men and women prefer different types of games and engage in different types of activities while gaming. Men play more action and simulation games, because they are more competition minded, but they are also more likely to get addicted to games (Huanhuan and Su, 2013). Women prefer logic and skill training games and participate in stereotypical female activities, such as shopping and dress up. While gaming, women are more likely to make friends, join groups, seek help and recruit new members. Many of these play patterns can be traced back to the disparate motives of men and women online. Overall, men online behavior is more goal-oriented, while women's activities are relationship and socialization focused (Gefen and Ridings, 2005).

Third, women's motives for playing games are similar to men's. Both genders enjoy the fantasy aspect of the games and like living through their virtual identities, both men and women seek relationships, diversion from daily lives and excitement. Thus, we find that women are interested in gaming experience and could be enticed to play as long as the games offer them the types of activities that they enjoy, incorporate social components and provide quick training for the required skills.

As we continue to rely more on games for education and training, developers need to ensure that the games are attractive for both men and women. Our findings suggest that women and men skills differ when it comes to gaming, which may further discourage women from participating. Therefore, learning and training games should be gender neutral and include components for motor and spatial skill training to initially help females with navigation and play. Further, educational gaming platforms should promote safe and comfortable social environment which does not alienate women with its stereotypical and sexist content (Fox and Tang, 2013).

Our findings also have implications for game developers and marketers. Since existing gaming culture is frequently a deterrent for women participation (Yee, 2006), developers can design games and worlds specifically for women or incorporate more social, family-oriented and visually appealing features into existing games. Finally, successfully leveraging online games for marketing requires more than establishing a commercial presence. Understanding gender preferences can help marketers to sharpen the precision in targeting their product and service ads or even develop particular product or service focused games for different population segments.

5. Conclusion

While a multitude of studies examine various aspects of gaming behavior, motives, outcomes and performance, few report gender differences. The ones that do, however, suggest that men and women gaming choices and behavior differ. We review and structured these insights to identify differences and similarities and recommend possible strategies to achieve gender parity in online gaming.

References

Becerra, E. P. and Stutts, M. A. 2008. “Ugly Duckling by Day, Super Model by Night: The Influence of Body Image on the Use of Virtual Worlds ,” *Journal of Virtual Worlds Research* (1:2), pp.1941-8477.

Chen, L. S.-L. 2010. “The impact of perceived risk, intangibility and consumer characteristics on online game playing,” *Computers in Human Behavior* (26:6), pp.1607–1613.

Choi, G., Chung, H. and Kim, Y. 2012. “Are Stereotypes Relative to Gender Usage Applicable to Virtual Worlds?,” *International Journal of Human–Computer Interaction* (28:6), pp. 399–405.

Chou, C. and Tsai, M.-J. 2007. “Gender differences in Taiwan high school students computer game playing,” *Computers in Human Behavior* (23:1), pp. 812–824.

Cohen, A. M. 2009. “Closing the Gender Gap in Online Gaming,” *THE FUTURIST* (43:6), pp. 10.

Ducheneaut, N., Wen, M.-H., D., Yee, N. and Wadley, G. 2009. “Body and Mind: A Study of Avatar Personalization in Three Virtual Worlds,” *New Media Experiences* (1), pp. 1151-1160.

Dunn, R. A. and Guadagno, R. E. 2012. “My avatar and me – Gender and personality predictors of avatar-self discrepancy,” *Computers in Human Behavior* (28:1), pp. 97-106.

Entertainment Software Association. 2013. “The 2013 Essential Facts About the Computer and Video Game Industry,” available at http://www.theesa.com/facts/pdfs/esa_ef_2013.pdf. Last accessed 02/26/2014.

Fox, J. and Tang, W. Y. 2013. “Sexism in online video games: The role of conformity to masculine norms and social dominance orientation,” *Computers in Human Behavior*, article in process.

Gefen, D. and Ridings, C. M. 2005. “If you Spoke as She Does, Sir, instead of the Way You Do: A Sociolinguistics Perspective of Gender Differences in Virtual Communities,” *SIGMIS Database* (36:2), pp. 78-92.

Guadagno, R. E., Muscanell, N. L., Okdie, B. M., Burk, N. M. and Ward, T.B. 2011. “Even in virtual environments women shop and men build: A social role perspective on Second Life,” *Computers in Human Behavior* (27:1), pp. 304-308.

Hainey, T., Connolly, T., Stansfield, M. and Boyle, E. 2011. “The differences in motivations of online game players and offline game players: A combined analysis of three studies at higher education level,” *Computers and Education* (57:1), pp.2197–2211.

Hassouneh, D. and Brengman, M. 2014. “A motivation-based typology of social virtual world users,” *Computers in Human Behavior* (33: April), pp. 330–338.

Hong, J.-C. and Hwang, M.-Y. 2012. “Gender Differences in Help-Seeking and Supportive Dialogue during on-Line Game,” *Procedia – Social and Behavioral Sciences* (64:9), pp. 342-351.

Hou, H.-T. 2012. “Exploring the behavioral patterns of learners in an educational massively multiple online role-playing game,” *Computers and Education* (58:4), pp. 1225–1233.

Iqbal, A., Kankaanranta, M. and Neittaanmäki, P. 2010. “Experiences and motivations of the young for participation in virtual worlds,” *Procedia – Social and Behavioral Sciences* (2:2), pp. 3190-3197.

Huanhuan, L. and Su, W. 2013. “The role of cognitive distortion in online game addiction among Chinese adolescents,” *Children and Youth Services Review* (35:9), pp. 1468–1475.

Kuo, W.-C., Wang, S.-T. and Yang, J.-C. 2012. “An empirical analysis of the playing time by different genders and ages in an MMORPG,” in *IEEE Computer Society*, Takamatsu, pp. 114-116.

Kuznekoff, J. H., and Rose, L. M. 2013. “Communication in multiplayer gaming: Examining player responses to gender cues,” *New Media & Society*, 15, pp. 541–556.

Lehdonvirta, M., Nagashima, Y., Lehdonvirta, V., and Baba, A. 2012. "The Stoic Male How Avatar Gender Affects Help-Seeking Behavior in an Online Game," *Games and Culture* (7:1), pp. 29-47.

Lever, J. 1976. “Sex Differences in the Games Children Play,” *Social Problems* (23:4), pp. 478-487.

Li, H., Liu, Y., Xu, X., and Heikkilä, J. 2013. “Please stay with me! An empirical investigation on hedonic IS continuance model for social network games,” in *Proceedings of the 34th International Conference on Information Systems*, Milano, Italy.

Liu, D., Li, X., and Santhanam, R. 2013. "Digital Games and Beyond: What Happens When Players Compete?," *MIS Quarterly*, (37:1), pp.111-124.

Lomanowska, A. M. and Guitton, M. J. 2013. “My avatar is pregnant! Representation of pregnancy, birth, and maternity in a virtual world,” *Computers in Human Behavior* (31), pp. 322-331.

Lucas, K., and Sherry, J. L. 2004. “Sex differences in video game play: a communication-based explanation,” *Communication Research* (31:5), pp. 499–523.

Quaiser-Pohl, C., Geiser, C. and Lehmann, W. 2005. “The relationship between computer-game preference, gender , and mental-rotation ability,” *Personality and Individual Differences* (40:3), pp. 609-619.

Schroeder, R. 2006. “Being there and the future of connected presence,” *Journal of Teleoperators and Virtual Environments* (15:4), pp. 438-454.

Tippett, W.J., Lee, J. H., Mraz, R., Zakzanis, K. K., Snyder, P. J., Black, S. E. and Graham, S. J. 2009. “Convergent Validity and Sex Differences in Healthy Elderly Adults for Performance on 3D Virtual Reality Navigation Learning and 2D Hidden Maze Tasks,” *Cyber Psychology & Behavior* (12:2), pp. 169-174.

Tlaukaa, M., Brolese, A., Pomeroy, D. and Hobbs, W. 2005. "Gender differences in spatial knowledge acquired through simulated exploration of a virtual shopping centre," *Journal of Environmental Psychology* (25:1), pp. 111-118.

van Reijmersdal, E.A., Jansz, J., Peters, O. and van Noort, G. 2013. "Why girls go pink: Game character identification and game-players' motivations," *Computers in Human Behavior* (29:6), pp.2640-2649.

Venkatesh, V. and Morris, M. G. 2000. "Why Don't Men Ever Stop to Ask for Directions? Gender, Social Influence, and Their Role in Technology Acceptance and Usage Behavior," *MIS Quarterly* (24:1), pp. 115-139.

Wang, E. S-T. 2013. "Perceived control and gender differences on the relationship between trialability and intent to play new online games," *Computers in Human Behavior* (30), pp. 315-320.

Webster, J. and Watson, R. T. 2002. "Analyzing the Past to Prepare for the Future: Writing a Literature Review," *MIS Quarterly* (26:2), pp. 13-23.

Williams, D., Ducheneaut, N., Xiong, L., Zhang, Y., Yee, N., and Nickell, E. 2006. "From Tree House to Barracks The Social Life of Guilds in World of Warcraft," *Games and culture* (1:4), pp. 338-361.

Williams, D., Yee, N. and Caplan, S.E. 2008. "Who plays, how much, and why? Debunking the stereotypical gamer profile," *Journal of Computer-Mediated Communication* (13:4), pp. 993-1018.

Yee, N. 2006. "Motivations for Play in Online Games," *Cyber Psychology & Behavior* (9:6), pp. 772-775.

Yee, N., Bailenson, J.N., Urbanek, M., Chang, F. and Merget, D. 2007. "The Unbearable Likeness of Being Digital: The Persistence of Nonverbal Social Norms in Online Virtual Environments," *Cyber Psychology & Behavior* (10:1), pp. 115-121.

Zhang, Y., Yu, X., Dang, Y. and Chen, H. 2010. "An Integrated Framework for Avatar Data Collection from the Virtual World: A Case Study in Second Life," *IEEE Intelligent Systems, (99), pp. 1.*

Zhou, Z., Jin X-L., Vogel, D.R., Fang, Y. and Chen, X. 2011. "Individual motivations and demographic differences in social virtual worlds uses: An exploratory investigation in Second Life," *International Journal of Information Management* (31:3), pp. 261-271.

ARTICLE 14:

GENDER DIFFERENCES IN ONLINE DATING: WHAT DO WE KNOW SO FAR? A SYSTEMATIC LITERATURE REVIEW[18]

Authors: Olga Abramova, Technische Universität Darmstadt, Darmstadt, Germany

Annika Baumann, Humboldt-Universität zu Berlin, Berlin, Germany

Hanna Krasnova, Universität Potsdam, Potsdam, Germany

Peter Buxmann, Technische Universität Darmstadt, Darmstadt, Germany

Proceedings: 49th Hawaii International Conference on System Sciences (HICSS'49)

Publication year: 2016

Abstract

With millions of users worldwide, online dating platforms strive to assert themselves as powerful tools to find dates and form romantic relationships. However, significant differences exist in male and female use of this mate-matching technology with respect to motivation, preferences, self-presentation, interaction and outcomes. While existing research has routinely reported on gender differences in online dating, these insights remain scattered across multiple studies. To gain a systematic insight into existing findings, in this study we conduct a meta-review of existing research. We find that evolutionary theory generally holds true in online dating: Users still follow natural stereotypes when it comes to choosing a mate online. Physical attractiveness is the key criteria for men; while women, being much more demanding, prioritize socio-economic attributes when choosing a male partner. Together, our structured findings offer a deeper insight into the underlying dynamics of gender differences in online dating.

1. Introduction

Online dating industry is enjoying a booming success: 11% of U.S. adults have already used a dating platform or an app, and a whopping 23% of users admit to having met their spouse or a long-term companion online (Smith and Duggan, 2013). As such, online dating represents a place where connections can be formed and dissolved quickly at little cost to both sides, offline social norms are less pronounced, and gender stereotypes can be mitigated by the initial anonymity of the dating partners (Žakelj, 2014). Making use of these affordances, both men and women readily embrace online dating channels in the search for a new companion, a short flirt, or even a long-term partnership. While statistics varies from country to country, and across different age groups, both men and women readily use popular dating platforms like Match.com, Yahoo Personals, eHarmony, and OkCupid, suggesting a strong interest of both parties in romantic interactions and connections (Sautter et al., 2010).

Nonetheless, their preferences, behaviors and choices are likely to differ (Buss, 1988). Indeed, in line with the evolutionary perspective on mating (Darwin, 1871), men and women exhibit distinct selection criteria when it comes to choosing their mating partners – differences that have far-reaching implications for both scholars of social behavior as well as system designers. However, while a wave of studies investigating various user-related aspects of online dating sites have captured a variety of gender-specific differences, these insights are scattered and do not provide a coherent picture of the state-of-the-art research available in this domain. To fill this gap, in this study we undertake a comprehensive review of existing empirical investigations to systematically summarize available knowledge in the area of gender differences in online dating.

This research is important for a number of reasons. On the theoretical side, our study will allow capturing the current research status quo, thereby helping to identify existing gaps open for future research. Furthermore, online-driven transformations in the mating behavior can be traced on the basis of our findings, allowing for better comparisons with the established knowledge from offline domain (e.g., Buss, 1989). From the managerial perspective, our study may empower platform providers in deciding on the gender-specific add-on features or special offers for the VIP platform areas common for such websites. On a more general level, by advancing knowledge in this domain this research may also contribute to a greater social good, since couples who meet via online channels have been shown be more satisfied and less likely to divorce, suggesting a favorable impact of online dating on the society at large (Cacioppo et al., 2013).

2. Theoretical Foundations

Social role (Eagly, 2013), self-construal (Cross and Madson, 1997) as well as evolutionary (Buss, 1988) theories have been often used to explain the differences in behavior and perceptions between men and women. In the dating context, particularly the evolutionary viewpoint is of critical importance, considering its focus on the choices of human species in the face of competition and search for limited reproduction-relevant opportunities. Originally formulated as a theory of sexual selection by (Darwin, 1871), this perspective suggests that reproductive success is a key evolutionary aspiration of human species, with both men and women striving to achieve the best possible outcome in this domain.

A distinction between intrasexual and intersexual selection is often made. Intrasexual selection implies competition among representatives of the same sex for a desired mating access. Here, competing agents are expected to produce signals that are viewed as desirable by the members of the opposite sex. At the same time, intersexual selection implies preferential choice exerted by members of one sex group with regard to the representatives of the other group (Buss, 1988). Conceptually representing two sides of the same phenomenon, the concepts of intra- and intersexual selection are closely related with the notion of parental investment (Trivers, 1972). According to this perspective, those who are expected to bear a higher parental investment in terms of nurturing and caring for potential offsprings are likely to be more selective with regard to their mating targets; at the same time, those who are less invested in the parental process will be less discriminative when choosing mating partners, striving to maximize "copulatory opportunities" (Buss, 1988, p. 617). However, they will also face greater competition to achieve reproductive access, and will have to correspond to and present selection-relevant attributes to the "choosing" party.

Since in many species these are the females who have to overtake the largest share of the parental investment, they are also more likely to be more selective with regard to their choice of male mates. These choices will be dictated by the male ability to compensate for his lack of parental investment or by his ability to provide it. Indeed, material resources, earning potential, social status, psychological support, protection, and such traits as ambition and industriousness have been consistently shown to play a role in the female choice of male partners in offline settings (Buss, 1989).

Nonetheless, since modern society often equates reproduction access with monogamy, men also face costly choices. Hence, they are likely to emphasize health, "good genes", physical attractiveness, youth and other "female" qualities that may appear important for the fulfillment of the female reproductive function (Buss, 1989; Buss, 1988).

So far, past research has provided empirical evidence for the existence of evolutionary-driven differences between male and female behaviors and perceptions in the offline context (Buss, 1989; Buss, 1988). At the same time, little systematic evidence exists on the gender differences in the modern context of online dating. Considering a growing independence of women and the rising emphasis on gender equality in the developed world (Inglehart and Norris, 2003), it might be possible that traditionally-assumed differences are no longer salient or at least undergo some degree of transformation. In the following, extant literature will be reviewed with regard to gender differences in online dating.

3. Methodology

Following the advice of Levy and Ellis (2006) and Webster and Watson (2002) we conducted a systematic literature review using the scientific databases ScienceDirect (154), EBSCOhost (211), Springer (791), Wiley Online Library (1091), Emerald Insights (47), JSTOR (205), ACM Digital Library (189), IEEE (97) and Google Scholar (12600)[19] in combination with the keywords {online dating OR digital dating OR dating website OR online mate OR internet dating OR internet romantic relationship OR online romance OR cyber flirting OR online love OR Match.com OR eHarmony} and {gender OR men OR women OR male OR female OR woman OR man OR sex differences}. We focused on English language sources, included only published articles and excluded books from our review. No other filters were applied. To be relevant, papers needed to have online dating in the focus of their research. The evaluation of relevance was based on the title and abstract. In the next step, all articles initially evaluated as relevant were checked for the presence of gender-related empirical results using the in-text search in combination with the gender-marked keywords stated above. Additionally, we conducted a backward and forward search to look for further relevant articles. This procedure resulted in 69 relevant articles published between the years 1995 to 2015. 73.19 % of them were published in journals, 23.19 % in conferences and the remaining two by the Pew Research Center (Madden and Lenhart, 2006; Smith and Duggan, 2013). The most popular publication outlets included such journals

[19] Numbers in brackets reflect the overall initial number of resultant papers.

as Computers in Human Behavior (8 studies), followed by Communication Research (3 studies). In terms of method, studies in our sample were based on surveys (27), interviews (7), experiments (7), descriptive analysis (26) or other types of statistical analyses (9). Around 40% of articles focused on gender-related issues; student samples were present in only a small share of all articles in our sample (10%).

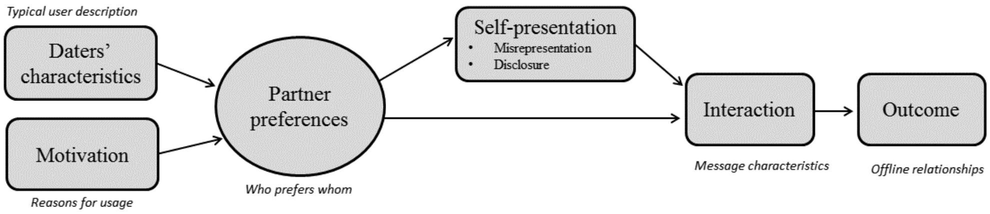

Figure 1. Process model of online dating.

In the following step, a total of 345 gender-relevant insights were derived using the in-text search, and then reviewed by two authors to identify the presence of the leading themes. Following this analysis we were able to elicit 7 different themes (see Table 1), that reflect a typical online dating process (see Figure 1). Next, two coders independently classified all insights into one of these themes: Cohen's Kappa (0.898) revealed a high level of agreement between the two coders (Landis and Koch, 1977).

Theme	Description	# of insights (share)
Daters' Characteristics	Characteristics of the user population.	43 (13.2%)
Motivation	Motivational patterns of users.	12 (3.7%)
Preferences	Preferences of users with regard to mating choices.	149 (45.9%)
Disclosure	Information shared on the profiles of users.	38 (11.7%)
Misreporting	Attributes misrepresented by users.	35 (10.8%)
Interaction	Dynamics of interaction between users via private messaging functionality.	46 (14.2%)
Outcome	Offline consequences resultant from online dating.	22 (6.8%)

Table 1. Themes in gender-relevant discourse on online dating.

4. Results

4.1. Daters' Characteristics

Since the experience of online dating revolves around the people who participate in it, insights centered on user characteristics were collected (see Table 2).

It appears that males are more active users of online dating sites: They use this service more (Fiore and Donath, 2005; Fiore et al., 2010; Gunter, 2008; Hitsch et al., 2010a; Hitsch et al., 2010b; Kisilevich and Last, 2010; Lin and Lindquist, 2013; Madden and Lenhart, 2006; Scharlott and Christ, 1995; Schmitz et al., 2011) and interact more on it (Gane, 2005; Hitsch et al., 2010a; Hitsch et al., 2010b; Lin and Lindquist, 2013; Valkenburg and Peter, 2007). One possible reason for this gender imbalance is that across numerous IT contexts men outnumber women (Liu, 2009). Additionally, male users exhibit a positive attitude towards online dating, valuing its efficiency to meet people (Liu, 2009; Madden and Lenhart, 2006), even though they might see those who use these service as desperate (Madden and Lenhart, 2006).

Sub-theme	Females	Males
Participa-tion	…use online dating more (Liu, 2009).	…use online dating more (Fiore and Donath, 2005; Fiore et al., 2010; Gunter, 2008; Hitsch et al., 2010a; Hitsch et al., 2010b; Kisilevich and Last, 2010; Lin and Lindquist, 2013; Madden and Lenhart, 2006; Scharlott and Christ, 1995; Schmitz et al., 2011).
	No gender differences (Sautter et al., 2010; Valkenburg and Peter, 2007).	
User Behavior		…view more profiles (Gane, 2005; Hitsch et al., 2010a; Hitsch et al., 2010b). …have longer account lifetimes (Lin and Lindquist, 2013). …visit online dating sites more often (Valkenburg and Peter, 2007).
	No gender differences in time spent on website (Lin and Lindquist, 2013).	
Attitude	…have more positive attitudes towards their online dating agency (Gunter, 2008).	…are more likely to have a positive attitude towards online dating in general (Liu, 2009; Madden and Lenhart, 2006).
Further Attributes	…report lower weight compared to the average (Hitsch et al., 2010a; Hitsch et al., 2010b). …report higher height compared to the average (Hitsch et al., 2010a; Hitsch et al., 2010b; Lin and Lindquist, 2013).	…report higher weight compared to the average (Hitsch et al., 2010a; Hitsch et al., 2010b). …report higher height compared to the average (Hitsch et al., 2010a; Hitsch et al., 2010b; Lin and Lindquist, 2013). …are better educated (Glasser et al., 2009; Groom and Pennebaker, 2005).

Table 2. Characteristics of online daters.

4.2. Motivation

Initial motives to engage in online dating are likely to play an important role in the subsequent process of mate selection. As such, sex differences in motivation are congruent with typical strategies of mating theory (see Table 3). When presented with a variety of opportunities, male users prefer short-term romantic relationships with low level of commitment (Clemens et al., 2015; Gunter, 2008; Scharlott and Christ, 1995). In contrast, female users claim to be driven by such a non-romantic reason as finding friends (Clemens et al., 2015; Gunter, 2008; Scharlott and Christ, 1995) or a potential marriage partner (Gunter, 2008), which, however, reveals inclination towards long-term relationships. Interestingly, these motivations are time-indifferent with studies dating 1995, 2008 and 2015 providing consistent results.

Motivation to use		Source
Male	Short-term (e.g., sex or date)	Clemens et al., 2015; Gunter, 2008; Scharlott and Christ, 1995
Female	Long-term (e.g., friendship)	Clemens et al., 2015; Gunter, 2008; Scharlott and Christ, 1995

Table 3. Motivation of online daters.

4.3. Preferences

The process of conscious mate-selection performed via online dating website implies a series of steps towards narrowing the pool of eligible candidates from many to one (Günaydin et al., 2013). In line with the differences in motivational patterns established above, our review suggests a relatively clear picture regarding male and female preferences for certain characteristics sought in a partner (see Table 4).

While females appear to value but be more tolerant towards the appearance of the potential partner (Hitsch et al., 2010b; Sritharan et al., 2010; Whitty, 2008), men do not hesitate to state exact body type preferences (Glasser et al., 2009; Hitsch et al., 2010a; Hitsch et al., 2010b) with thin and toned body types being most desired (Glasser et al., 2009). Indeed, physical attractiveness of a female appears to be a decisive criterion for male online daters (Alterovitz and Mendelsohn, 2009; Bak, 2010; Hitsch et al., 2010a; Hitsch et al., 2010b; Lee et al., 2008; McWilliams and Barrett, 2014; Scharlott and Christ, 1995; Whitty, 2008), corresponding to their search for female reproduction advantage.

Characteristics of a partner	Direction of preference compared to the self	Preferences of:	
		Male	**Female**
Education	Up (Higher)		Hitsch et al., 2010a; Schmitz et al., 2009; Xia et al., 2014
	Homophily	Hitsch et al., 2010a; Hitsch et al., 2010b; Schmitz et al., 2009	Hitsch et al., 2010a; Hitsch et al., 2010b; Schmitz et al., 2009; Skopek et al., 2011
	Down (Lower)	Hitsch et al., 2010a; Kreager et al., 2014; Skopek et al., 2010	
	Importance		Schmitz et al., 2009; Xia et al., 2014
Age	Up (Higher)		Alterovitz and Mendelsohn, 2009; Burrows, 2013; Fiore et al., 2010; Hitsch et al., 2010b; Kaufman and Phua, 2003; Kreager et al., 2014; Skopek et al., 2011; Xia et al., 2014
	Homophily	Skopek et al., 2011	Skopek et al., 2011
	Down (Lower)	Alterovitz and Mendelsohn, 2009; Burrows, 2013; Fiore et al., 2010; Hitsch et al., 2010b; Kaufman and Phua, 2003; Kreager et al., 2014; Skopek et al., 2011; Xia et al., 2014	Alterovitz and Mendelsohn, 2009; Burrows, 2013; Fiore et al., 2010 Skopek et al., 2011; Xia et al., 2014
Height	Up (Higher)		Hitsch et al., 2010a; Hitsch et al., 2010b; Kreager et al., 2014; Salska et al., 2008
	Down (Lower)	Hitsch et al., 2010a; Hitsch et al., 2010b; Salska et al., 2008	
Socio-economic status (income and occupation)	Up (Higher)	Hitsch et al., 2010b; Whitty, 2008	Anderson and Klofstad, 2012; Hitsch et al., 2010b; Ong and Wang, 2015; Scharlott and Christ, 1995; Whitty, 2008; Xia et al., 2014
	Down or no strong preference	Anderson and Klofstad, 2012; Xia et al., 2014	
	Importance		Anderson and Klofstad, 2012; Hitsch et al., 2010a; Hitsch et al., 2010b; McWilliams and Barrett, 2014; Whitty, 2008; Xia et al., 2014
Physical attractiveness	Body type preference	Glasser et al., 2009; Hitsch et al., 2010a; Hitsch et al., 2010b	
	Importance	Alterovitz and Mendelsohn, 2009 ; Bak, 2010; Hitsch et al., 2010a; Hitsch et al., 2010b; Lee et al., 2008; McWilliams and Barrett, 2014; Scharlott and Christ, 1995; Whitty, 2008	Hitsch et al., 2010b; Sritharan et al., 2010; Whitty, 2008
Preference for profile features of the opposite sex	Photos	Bak, 2010; Fiore and Donath, 2005; Kreager et al., 2014; Xia et al., 2014	Fiore and Donath, 2005; Xia et al., 2014
	Description	Fiore and Donath, 2005	Kreager et al., 2014

Table 4. Patterns of partner preferences of online daters.

With respect to age criterion there is a clear pattern for men to look for a younger (Alterovitz and Mendelsohn, 2009; Burrows, 2013; Fiore et al., 2010; Hitsch et al., 2010b; Kaufman and Phua, 2003; Kreager et al., 2014; Skopek et al., 2011; Xia et al., 2014) or at least a same-age partner (Skopek et al., 2011). Moreover, these preferences are invariant to the age of a man.

Quite on the contrary, female daters are better predisposed towards older male candidates (Alterovitz and Mendelsohn, 2009; Burrows, 2013; Fiore et al., 2010; Hitsch et al., 2010b; Kaufman and Phua, 2003; Kreager et al., 2014; Skopek et al., 2011; Xia et al., 2014). A more detailed investigation suggests that female age preference represents an inverted U-shape function of her own age. Starting with a strict preference for older partners, women broaden their preferred age ranges as they get older and show higher inclination towards homophily when they reach 25 years of age. However, aging women increasingly prefer younger partners (Fiore et al., 2010).

Recent research argues for the derivative nature of age choice hypothesizing that preferences for "women's age are (partially) a function of male preferences for physical attractiveness, whereas female preferences for men's age are (partially) a function of female preferences for male socio-economic status" (Skopek et al., 2011, p. 273). In the modern society that values fitness and youth, youthful look is one of the key attributes of physical attractiveness. Coupled with the biological fact that female fertility is affected by age stronger than male fertility, this warrants the age choice of men (Buss, 1988; Darwin, 1871). At the same time, females strongly prioritize socio-economic status (Anderson and Klofstad, 2012; Hitsch et al., 2010a; Hitsch et al., 2010b; McWilliams and Barrett, 2014; Whitty, 2008; Xia et al., 2014) when choosing a male partner, and, therefore, are more likely to prefer older and, hence, more financially mature male targets.

All in all, it is evident that female mating choice is congruent with the parental investment theory (Trivers, 1972). Women are pickier in specifying the type of partner they are looking for (Feliciano et al., 2009; Fiore et al., 2010; Glasser et al., 2009; Lee et al., 2008; Toma and Hancock, 2010; Xia et al., 2014). The fact that family's material well-being may depend on male income (Schmitz et al., 2009) may explain strong preference of women to date wealthier men (Anderson and Klofstad, 2012; Hitsch et al., 2010b; Ong and Wang, 2015; Scharlott and Christ, 1995; Whitty, 2008; Xia et al., 2014). According to our review, this also holds true for high earning women (Ong and Wang, 2015). At the same time, men are more open and are ready to become acquainted with women with lower income (Anderson and Klofstad, 2012; Xia et al., 2014). However, in general, both men and women prefer high-income partners over low-income partners (Hitsch et al., 2010b; Whitty, 2008), which can be explained as an attempt to avoid dating for mercenary ends.

Further, a well-established positive relationship between socio-economic status and academic achievements (Caro et al., 2009) explains the fact that educational preferences follow the same gender patterns as socio-economic status, and are much more critical for women (Schmitz et al., 2009; Xia et al., 2014). Higher academic degree of a man attracts women (Hitsch et al., 2010a; Schmitz et al., 2009; Xia et al., 2014), while educational homophily is considered to be a good choice for both women (Hitsch et al., 2010a; Hitsch et al., 2010b; Schmitz et al., 2009; Skopek et al., 2010) and men (Hitsch et al., 2010a; Hitsch et al., 2010b; Schmitz et al., 2009).

All in all, men are much less demanding with respect to their mate's education and willingly contact women with a lower academic degree (Hitsch et al., 2010a; Kreager et al., 2014; Skopek et al., 2010). However, men are not attracted by women's intelligence when it surpasses their own (Hitsch et al., 2010a).

Online daters' preferences for height follow "male-taller" norm (Hitsch et al., 2010a; Hitsch et al., 2010b; Kreager et al., 2014; Salska et al., 2008) for both cases, with preferences from the female side being more pronounced (Salska et al., 2008). Tall men and short women, however, are more tolerant to the disparity in height, thereby maximizing their dating pool. This is in contrast to tall women and short men who try to adhere to socially recognizable standard (Salska et al., 2008).

Finally, men and women also have certain preferences when it comes to the information members of the opposite sex provide. While all daters who posted more photos have a greater chance to convince potential partners in their own attractiveness (Fiore and Donath, 2005; Xia et al., 2014), posting photos is especially relevant for the dating success of women. For them, the number of received messages is positively related to the number of photos they post (Bak, 2010; Fiore and Donath, 2005; Kreager et al., 2014; Xia et al., 2014), once again indirectly proving the importance of physical attractiveness for men. In contrast, women prefer men to post longer self-descriptions (Xia et al., 2014) and perceive the candidate as more credible when rich media, such as video or audio, is used (Lee et al., 2008).

4.4. Disclosure

In order to allow for a match, both men and women present themselves to other participants of the online dating community, which implies a certain degree of disclosure (see Table 5).

Degree of disclosure		**Source**
Male	Disclose more information about themselves	Günaydin et al., 2013 Gunter, 2008 Lin and Lindquist, 2013
	Reveal more homogeneous information	Kisilevich and Last, 2010
Female	Provide more heterogeneous information and are more creative	Kisilevich and Last, 2010
Type of information more likely to be disclosed		**Source**
Male	Status-related information (income and occupation)	Alterovitz and Mendelsohn, 2009 Anderson and Klofstad, 2012 Fiore et al., 2010 Groom and Pennebaker, 2005 Lin and Lindquist, 2013 McWilliams and Barrett, 2014
	Phone numbers	Gane, 2005
	Photos	Kisilevich and Last, 2010
	Sex and cars	Kisilevich and Last, 2010
Female	Information about kids	Kisilevich and Last, 2010 Lin and Lindquist, 2013
	Desired age of a partner	Kisilevich and Last, 2010
	Photos	Whitty, 2008 Xia et al., 2014
	Interests	Whitty, 2008
	Home and sex	Fiore et al., 2010 Groom and Pennebaker, 2005
Type of information more likely to be disclosed		**Source**
Male	Typically describe themselves as average or athletic and fit	Glasser et al., 2009 Lin and Lindquist, 2013
	Use more numbers and social words in texts	Groom and Pennebaker, 2005
Female	Typically describe themselves as small or large and overweight	Glasser et al., 2009 Lin and Lindquist, 2013
	Use longer texts for self-description	Fiore et al., 2010 Groom and Pennebaker, 2005
	Use more positive emotion words, spatial words and personal pronouns in texts	Fiore et al., 2010 Groom and Pennebaker, 2005
Both	Use common and tentative words	Nagarajan and Hearst, 2009

Table 5. Disclosure patterns of online daters.

It is observed that male daters disclose more than their female counterparts (Gibbs et al., 2011; Gibbs et al, 2011; Gunter, 2008; Lin and Lindquist, 2013), even though their profiles are of rather standard, homogenous character, with a restricted range of information they choose to provide (Kisilevich and Last, 2010). In line with the mating theory, demonstrating resources that are highly desired by members of the opposite sex, men tend to disclose status-related information like income and occupation (Alterovitz and Mendelsohn, 2009; Anderson and Klofstad, 2012; Fiore et al., 2010; Groom and Pennebaker, 2005; Lin and Lindquist, 2013; McWilliams and Barrett, 2014) or cars (Kisilevich and Last, 2010). In the hope of moving the interaction to a more personal level, they are ready to provide photos (Kisilevich and Last, 2010), phone numbers (Gane, 2005) and sex-related information (Kisilevich and Last, 2010).

At the same time, women are more creative and multifarious in their self-presentations (Kisilevich and Last, 2010). They are more likely to provide information about their children (Kisilevich and Last, 2010; Lin and Lindquist, 2013), interests (Whitty, 2008) as well as home and sex (Fiore et al., 2010; Groom and Pennebaker, 2005). Understanding the importance of their physical attractiveness for the mating success, women readily upload more photos than men (Whitty, 2008; Xia et al., 2014).

Textual analysis of the information provided in the "About me" section shows typical gender patterns, with men using more numerals and references to other people (Groom and Pennebaker, 2005) and women using personal pronouns, positive emotion and spatial words as well as writing longer self-descriptions in general (Fiore et al., 2010; Groom and Pennebaker, 2005). However, no differences are observed in the use of frequent and tentative words (Hall et al., 2010). All in all, the patterns of disclosure follow predictions of the evolutionary theory described above (Buss, 1988; Darwin, 1871).

4.5. Misrepresentation

To achieve better matches, online daters are tempted to misrepresent certain desired attributes (Landolt et al., 1995) (see Table 6). To prevent this, participants are encouraged to formally report the presence of falsified information through feedback mechanisms available on some platforms.

Both men and women report that they have faced instances of misreporting on online dating sites (Smith and Duggan, 2013) suggesting that this behavioral tendency is rather common (Lo et al, 2013; Toma et al., 2008). However, different information is misrepresented by female and male daters (see Table 6). Aware of the importance of their physical attractiveness to men, females are more likely to use enhanced photographic material (Hancock and Toma, 2009; Lo et al, 2013; Schmitz et al., 2013; Toma and Hancock, 2010; Whitty, 2008), and underreport their weight (Close and Zinkhan, 2004; Hall et al., 2010; Hancock et al., 2007; Toma et al., 2008) and age (Close and Zinkhan, 2004) (even though the latter is also common for men (Hall et al., 2010)). This way, female users are trying to advance themselves in comparison to other female contenders, rank higher in search listings, and, thereby, achieve better matches (Buss, 1988).

In contrast, men tend to rather emphasize their personal interests and assets (Hall et al., 2010) to gain a better hierarchical position in the competitive environment of online dating. This signaling behavior allows them access to a larger pool of females, who are generally seeking rather resource-rich males (Trivers, 1972).

Since height is an attribute often psychologically associated with strength and status (Buss, 1988) and both short and tall women prefer taller men (Hitsch et al., 2010a; Hitsch et al., 2010b; Kreager et al., 2014; Salska et al., 2008), male users also have the tendency to overstate this characteristic on their profile (Hancock et al., 2007; Hancock et al., 2007; Schmitz et al., 2013; Toma et al., 2008; Whitty, 2008). Furthermore, men have been found to misrepresent their current relationship status as well as the goals they want to achieve when using online dating services (Hall et al., 2010; Schmitz et al., 2013; Whitty, 2008). Possibly, they might do so to adapt their short-term focus to a rather long-term one of females (Rhodes et al., 2005), which is in line with the evolutionary theory (Buss, 1988; Darwin, 1871), since females have to invest more resources into the parental process (Trivers, 1972).

Information on:	**Females are more likely to misrepresent:**	**Males are more likely to misrepresent:**
Age	…their age (Close and Zinkhan, 2004).	…their age (Hall et al., 2010).
	No gender differences (Hancock et al., 2007).	
Height		…their height (Hancock et al., 2007; Schmitz et al., 2013; Toma et al., 2008; Whitty, 2008).
Physical Attractiveness	…their physical attractiveness (Hancock and Toma, 2009; Lo et al, 2013; Schmitz et al., 2013; Toma and Hancock, 2010; Whitty, 2008).	…their physical attractiveness (Gane, 2005).
Relationship		...their relationships status (Schmitz et al., 2013; Whitty, 2008) and goals (Hall et al., 2010; Schmitz et al., 2013).
Weight	…their weight (Close and Zinkhan, 2004; Hall et al., 2010; Hancock et al., 2007; Toma et al., 2008).	…their weight (Toma et al., 2008).
General	…themselves (Gane, 2005).	…themselves (Guadagno et al., 2012).

Table 6. Misrepresentation patterns of online daters.

4.6. Interaction

In terms of resulting interaction (see Table 7), there is a strong agreement in the literature that females receive more contacts by males who readily initiate a starting conversation (Bapna et al., 2013; Fiore and Donath, 2005; Fiore et al., 2010; Gane, 2005; Hitsch et al., 2010b; Kreager et al., 2014; Lewis, 2013; Scharlott and Christ, 1995; Xia et al., 2014). Moreover, functionality-enabled ability to see who visited one' profile is particularly encouraging for men (e.g. as offered on Match.com, eHarmony, Parship, OkCupid, and others), who are more likely to use this feature to send messages to females who visited their profile (Bapna et al., 2013). In line with the above, males also receive significantly fewer replies and messages in general (Bapna et al., 2013; Fiore and Donath, 2005; Fiore et al., 2010; Hitsch et al., 2010a; Kreager et al., 2014; Lin and Lindquist, 2013), whereas females can expect a lot of reciprocation (Fiore and Donath, 2005; Fiore et al., 2010; Hitsch et al., 2010a; Kreager et al., 2014; Scharlott and Christ, 1995; Xia et al., 2014). In their interactions, women tend to send more general messages (Xia et al., 2014) as well as are more likely to carry on the communication (Scharlott and Christ, 1995). Together, this suggests that males try to make use of the opportunity to have access to multiple females and are satisfied with a superficial character of such contacts. In contrast, women are rather picky in their decision of who might be their potential date (Arnold and Duvall, 1994). Interacting with fewer male users, women show interest in creating more intimate and intensive conversations (Scharlott and Christ, 1995).

Some characteristics, such as, for example, attractiveness or using only few self-references seem to increase the likelihood to receive a reply for both men and women (Schoendienst and Dang-Xuan, 2011). Additionally, explicitly stated dating preferences (Fiore et al., 2010) and a sexually-related talk (Schoendienst and Dang-Xuan, 2011) enhance the chances of reciprocation for female users. At the same time, lengthy messages enhance the chances for men to get a reply (Schoendienst and Dang-Xuan, 2011).

Types of interaction	Females are more likely to:	Males are more likely to:
Initiation	...receive more initial messages (Fiore et al., 2010; Gane, 2005).	…initiate contact (Bapna et al., 2013; Fiore and Donath, 2005; Fiore et al., 2010; Hitsch et al., 2010b; Kreager et al., 2014; Lewis, 2013; Scharlott and Christ, 1995; Xia et al., 2014). …receive more initiations relative to profile views (Gane, 2005).
Reciprocation	...receive a reply (Fiore and Donath, 2005; Fiore et al., 2010; Hitsch et al., 2010a; Kreager et al., 2014; Xia et al., 2014).	...not receive a reply (Bapna et al., 2013; Fiore and Donath, 2005; Fiore et al., 2010; Hitsch et al., 2010a).
General	...receive more messages (Kreager et al., 2014; Scharlott and Christ, 1995; Xia et al., 2014). ...send more messages (Xia et al., 2014). ...carry on the interaction (Scharlott and Christ, 1995).	...send more messages (Lin and Lindquist, 2013). ...receive fewer messages (Kreager et al., 2014; Lin and Lindquist, 2013). ...participate in more communications (Fiore and Donath, 2005).

Table 7. Interaction patterns of online daters.

4.7. Outcome

In the final step of the online dating process (see Figure 1) a shift to the offline environment might take place. It appears that females are rather reluctant to meet other users face-to-face since they need more computer-mediated interaction compared to males before an actual meeting offline (Gane, 2005). This might be connected to the circumstance that females are more likely to experience negative interactions on online dating sites (Smith and Duggan, 2013), which is also supported by the evidence that females are more likely to tell others about their plan to meet with another user in the offline setting (Blackhart et al., 2014). Even though one study reports higher first meeting rates for females (Gibbs et al., 2006), there is more evidence that both men and women tend to have a similar amount of first-date experiences offline using online dating platforms (Gunter, 2008; Smith and Duggan, 2013).

The first face-to-face meeting is the point where the fit with a potential partner is evaluated: Here, females have higher drop-out rates in terms of their subsequent evaluation of their dating partner (Norton et al., 2007). Again, this might suggest that men tend to focus on quantity, whereas females rather emphasize the "quality" of their dating partners (Rhodes et al., 2005; Trivers, 1972). Overall, studies report contradictory findings that either more females (Lever et al., 2008; Scharlott and Christ, 1995), more males (Cacioppo et al., 2013; Gunter, 2008) or both (Gunter, 2008; Smith and Duggan, 2013) have experienced a positive outcome in terms of various dating goals (e.g., long-term relationships or sexual relationship, with men reporting more sexual relationships (Gunter, 2008)). Together, however, this evidence suggests that using online dating services can be beneficial for both, even though more research is needed to gain a better understanding of this dynamics.

5. Concluding Remarks

In a delicate IT-driven business of online dating, providers are becoming increasingly attentive to how users make their choices. Understanding behavioral patterns enables providers to select relevant offers, thereby helping to increase the matching rate – one of the main goals of these platforms. Responding to this demand, this study provides an exhaustive summary of gender differences in behavior and perceptions of online daters. By focusing on heterosexual dating process, our findings reveal how gender intersects with daters' characteristics, motivation, preferences, disclosure, misrepresentation, interaction and offline outcomes. We analyze singles' online dating behavior in line with the evolutionary approach.

We observe that men are more active on online dating platforms. They are less choosy about partners and are more likely to be motivated by short-term romantic pleasure. While male online daters are attracted by physical appearance of a potential mate, female daters base their choices on male breadwinning abilities and give preference to socio-economic characteristics (income, occupation and education) over physical attractiveness. Although men disclose more readily, women lead in creativity and variety of information provided. However, both males and females are caught misrepresenting some of their information when creating their profiles. For example, digital enhancement of physical attractiveness is rather characteristic for female daters. At the same time, male users are more likely to falsify their relationships status and goals. Interacting on online dating platforms each party follows its conventional role: Men initiate more contacts, giving women a choice to reciprocate the attention and carry on the interaction. Regarding the outcome of online dating, gender differences remain unclear and offer an interesting venue for future research. Our study has several limitations: race-related and homosexual preferences were not in the scope of the current analysis. Moreover, cultural differences (Darwin, 1871) were not considered, thus paving the way for further investigations.

References

Alterovitz, S. S. R. and Mendelsohn, G. A., "Partner preferences across the life span: online dating by older adults", Psychology and aging, 24(2), 2009, p. 513.

Anderson, R. C., and Klofstad, C. A. "For love or money? The influence of personal resources and environmental resource pressures on human mate preferences", Ethology, 118(9), 2012, pp. 841-849.

Arnold, S. J. and Duvall, D. "Animal mating systems: a synthesis based on selection theory," American Naturalist, 1994, pp. 317-348.

Bak, P. "Sex differences in the attractiveness halo effect in the online dating environment", Journal of Business and Media Psychology, 1, 2010, pp. 1-7.

Bapna, R., Ramaprasad, J., Shmueli, G. and Umyarov, A., "One-way mirrors and weak-signaling in online dating: A randomized field experiment", International Conference on Information Systems (ICIS 2013): Reshaping Society Through Information Systems Design, 3, 2013, pp. 2748-2762.

Blackhart, G. C., Fitzpatrick, J., and Williamson, J., "Dispositional factors predicting use of online dating sites and behaviors related to online dating", Computers in Human Behavior, 33, 2014, pp. 113-118.

Burrows, K., "Age preferences in dating advertisements by homosexuals and heterosexuals: From sociobiological to sociological explanations", Archives of Sexual Behavior, 42(2), 2013, pp. 203-211.

Buss, D. M., "Sex differences in human mate preferences: Evolutionary hypotheses tested in 37 cultures", Behavioral and Brain Sciences, 12, 1989, pp. 1-49.

Buss, D.M., "The evolution of human intrasexual competition: Tactics of mate attraction", Journal of Personality and Social Psychology, 54(4), 1988, pp. 616–628.

Cacioppo, J. T., Cacioppo, S., Gonzaga, G. C., Ogburn, E. L. and VanderWeele, T. J., "Marital satisfaction and break-ups differ across on-line and off-line meeting venues", Proceedings of the National Academy of Sciences, 110(25), 2013, pp. 10135-10140.

Caro, D. H., McDonald, J. T. and Willms, J. D. "Socio-economic status and academic achievement trajectories from childhood to adolescence," Canadian Journal of Education/ Revue canadienne de l'éducation, 32(3), 2009, pp. 558-590.

Clemens, C., Atkin, D., and Krishnan, A., "The influence of biological and personality traits on gratifications obtained through online dating websites", Computers in Human Behavior, 49, 2015, pp. 120-129.

Close, A. G., and Zinkhan, G., "Romance and the Internet: The E-Mergence of Edating", Advances in Consumer Research, 31, 2004, pp. 153-157.

Cross, S. E. and Madson, L.,"Models of the Self: Self-Construals and Gender", Psychological Bulletin, 122(1), 1997, pp. 5-37.

Darwin, C., The descent of man and selection in relation to sex, J. Murray, London, 1871.

Eagly, A. H., Sex differences in social behavior: A social-role interpretation, Hillsdale, NJ: Erlbaum, Psychology Press, 2013.

Feliciano, C., Robnett, B., and Komaie, G., "Gendered racial exclusion among white internet daters", Social Science Research, 38(1), 2009, pp. 39-54.

Fiore, A. T., and Donath, J. S., "Homophily in online dating: when do you like someone like yourself?", In CHI'05 Extended Abstracts on Human Factors in Computing Systems, ACM, 2005, pp. 1371-1374.

Fiore, A. T., Taylor, L. S., Zhong, X., Mendelsohn, G. A. and Cheshire, C., "Who's right and who writes: People, profiles, contacts, and replies in online dating", In hicss, IEEE, 2010, pp. 1-10.

Gane, M., "Gender differences in online dating: An investigation of vulnerability and deception", Annual meeting of the American Sociological Association, 2005.

Gibbs, J. L., Ellison, N. B. and Heino, R. D., "Self-presentation in online personals the role of anticipated future interaction, self-disclosure, and perceived success in Internet dating" Communication Research, 33(2), 2006, pp. 152-177.

Gibbs, J. L., Ellison, N. B. and Lai, C. H., "First comes love, then comes Google: An investigation of uncertainty reduction strategies and self-disclosure in online dating" Communication Research, 2011, 0093650210377091.

Glasser, C. L., Robnett, B. and Feliciano, C., "Internet daters' body type preferences: Race–ethnic and gender differences", Sex roles, 61(1-2), 2009, pp. 14-33.

Groom, C. J. and Pennebaker, J. W., "The language of love: Sex, sexual orientation, and language use in online personal advertisements", Sex Roles, 52(7-8), 2005, pp. 447-461.

Guadagno, R. E., Okdie, B. M. and Kruse, S. A., "Dating deception: Gender, online dating, and exaggerated self-presentation", Computers in Human Behavior, 28(2), 2012, pp. 642-647.

Günaydin, G., Selcuk, E., and Hazan, C. "Finding the one: A process model of human mate selection," Human bonding, 2013, pp. 103-131.

Gunter, B., "Internet dating: A British survey", In Aslib Proceedings 60(2), Emerald Group Publishing Limited, 2008, pp. 88-98.

Hall, J. A., Park, N., Song, H. and Cody, M. J., "Strategic misrepresentation in online dating: The effects of gender, self-monitoring, and personality traits", Journal of Social and Personal Relationships, 27(1), 2010, pp. 117-135.

Hancock, J. T. and Toma, C. L., "Putting your best face forward: The accuracy of online dating photographs", Journal of Communication, 59(2), 2009, pp. 367-386.

Hancock, J. T., Toma, C. and Ellison, N., "The truth about lying in online dating profiles", In Proceedings of the SIGCHI conference on Human factors in computing systems, ACM, 2007, pp. 449-452.

Hitsch, G. J., Hortaçsu, A. and Ariely, D., "Matching and sorting in online dating" The American Economic Review, 2010a, pp. 130-163.

Hitsch, G. J., Hortaçsu, A., and Ariely, D., "What makes you click? - Mate preferences in online dating", Quantitative marketing and Economics, 8(4), 2010b, pp. 393-427.

Inglehart, R. and Norris, P. Rising tide: Gender equality and cultural change around the world. Cambridge University Press. 2003

Kaufman, G. and Phua, V. C., "Is ageism alive in date selection among men? Age requests among gay and straight men in Internet personal ads", The Journal of Men's Studies, 11(2), 2003, pp. 225-235.

Kisilevich, S. and Last, M., Exploring gender differences in member profiles of an online dating site across 35 countries, Springer-Verlag, Berlin, Heidelberg pp. 57–78, 2010.

Kreager, D. A., Cavanagh, S. E., Yen, J. and Yu, M., "Where Have All the Good Men Gone?" Gendered Interactions in Online Dating, Journal of Marriage and Family, 76(2), 2014, pp. 387-410.

Landis, J. R. and Koch, G. G. "The measurement of observer agreement for categorical data," Biometrics 33 (1), 1977, pp. 159-174.

Landolt, M. A., Lalumière, M. L., & Quinsey, V. L. (1995). "Sex differences in intra-sex variations in human mating tactics: An evolutionary approach," Ethology and Sociobiology, 16(1), 1995, pp. 3-23.

Lee, L., Loewenstein, G., Ariely, D., Hong, J. and Young, J., "If I'm not hot, are you hot or not? Physical-attractiveness evaluations and dating preferences as a function of one's own attractiveness", Psychological Science, 19(7), 2008, pp. 669-677.

Lever, J., Grov, C., Royce, T. and Gillespie, B. J., "Searching for love in all the "write" places: Exploring Internet personals use by sexual orientation, gender, and age", International Journal of Sexual Health, 20(4), 2008, pp. 233-246.

Levy, Y., and Ellis, T. J. 2006. "A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research," Informing Science 9 (1), 2006, pp. 181-212

Lewis, K., "The limits of racial prejudice", Proceedings of the National Academy of Sciences, 110(47), 2013, pp. 18814-18819.

Lin, K. H. and Lundquist, J., "Mate Selection in Cyberspace: The Intersection of Race, Gender, and Education", American Journal of Sociology, 119(1), 2013, pp. 183-215.

Liu, X., "Sociological Examination of People's Attitude Towards Online Dating" The annual meeting of the American Sociological Association Annual Meeting, San Francisco, 2009.

Lo, S. K., Hsieh, A. Y. and Chiu, Y. P., "Contradictory deceptive behavior in online dating" Computers in Human Behavior, 29(4), 2013, pp. 1755-1762.

Madden, M. and Lenhart, A., “Online dating: Americans who are seeking romance use the Internet to help them in their search, but there is still widespread public concern about the safety of online dating”, Pew Internet & American Life Project, 2006.

McWilliams, S. and Barrett, A. E., “Online dating in middle and later life gendered expectations and experiences”, Journal of Family Issues, 35(3), 2014, pp. 411-436.

Nagarajan, M. and Hearst, M. A., “An Examination of Language Use in Online Dating Profiles”, In ICWSM, 2009.

Norton, M. I., Frost, J. H. and Ariely, D., “Less is more: the lure of ambiguity, or why familiarity breeds contempt”, Journal of personality and social psychology, 92(1), 2007, p. 97.

Ong, D. and Wang, J., “Income attraction: An online dating field experiment”, Journal of Economic Behavior & Organization, 111, 2015, pp. 13-22.

Rhodes, G., Simmons, L. W. and Peters, M. “Attractiveness and sexual behavior: Does attractiveness enhance mating success?. Evolution and human behavior,” 26(2), 2005, pp. 186-201.

Salska, I., Frederick, D. A., Pawlowski, B., Reilly, A. H., Laird, K. T. and Rudd, N. A., “Conditional mate preferences: Factors influencing preferences for height”, Personality and Individual Differences, 44(1), 2008, pp. 203-215.

Sautter, J. M., Tippett, R. M. and Morgan, S. P., “The social demography of internet dating in the United States*”, Social Science Quarterly, 91(2), 2010, pp. 554-575.

Scharlott, B. W. and Christ, W. G., “Overcoming relationship-initiation barriers: The impact of a computer-dating system on sex role, shyness, and appearance inhibitions”, Computers in Human Behavior, 11(2), 1995, pp. 191-204.

Schmitz, A., Sachse-Thürer, S., Zillmann, D. and Blossfeld, H. P., “Myths and facts about online mate choice Contemporary beliefs and empirical findings”, Zeitschrift für Familienforschung - Journal of Family Research, 23(3), 2011.

Schmitz, A., Skopek, J., Schulz, F., Klein, D. and Blossfeld, H. P., “Indicating mate preferences by mixing survey and process-generated data. The case of attitudes and behaviour in online mate search”, Historical Social Research/Historische Sozialforschung, 2009, pp. 77-93.

Schmitz, A., Zillmann, D. and Blossfeld, H., “Do women pick up lies before men? The association between gender, deception patterns, and detection modes in online dating”, Online Journal of Communication and Media Technologies, 3(3), 2013, pp. 52-73

Schoendienst, V. and Dang-Xuan, L., “The Role of Linguistic Properties in Online Dating Communication-A Large-Scale Study of Contact Initiation Messages”, In PACIS, 2011, p. 169.

Skopek, J., Schmitz, A. and Blossfeld, H. P., “The gendered dynamics of age preferences–Empirical evidence from online dating”, Zeitschrift für Familienforschung - Journal of Family Research, 23(3), 2011.

Skopek, J., Schulz, F. and Blossfeld, H. P., “Who contacts whom? Educational homophily in online mate selection”, European Sociological Review, 2010, jcp068.

Smith, A. and Duggan, M., “Online dating & relationships”, Pew Internet & American Life Project, 2013.

Sritharan, R., Heilpern, K., Wilbur, C. J. and Gawronski, B., ”I think I like you: Spontaneous and deliberate evaluations of potential romantic partners in an online dating context”, European Journal of Social Psychology, 40(6), 2010, pp. 1062-1077.

Toma, C. L. and Hancock, J. T., “Looks and lies: The role of physical attractiveness in online dating self-presentation and deception”, Communication Research, 37(3), 2010, pp. 335-351.

Toma, C. L., Hancock, J. T. and Ellison, N. B., “Separating fact from fiction: An examination of deceptive self-presentation in online dating profiles”, Personality and Social Psychology Bulletin, 34(8), 2008, pp. 1023-1036.

Trivers, R.L., Parental investment and sexual selection. Sexual selection and the descent of man, Aldine de Gruyter, New York, pp. 136-179, 1972.

Valkenburg, P. M. and Peter, J., “Who visits online dating sites? Exploring some characteristics of online daters”, CyberPsychology & Behavior, 10(6), 2007, pp. 849-852.

Webster, J., and Watson, R. T. “Analyzing the past to prepare for the future: Writing a Literature Review,” MIS Quarterly, 26(2), 2002, pp. 13-23.

Whitty, M. T., “Revealing the ‘real’me, searching for the ‘actual’you: Presentations of self on an internet dating site”, Computers in Human Behavior, 24(4), 2008, pp.1707-1723.

Xia, P., Jiang, H., Wang, X., Chen, C. and Liu, B., “Predicting user replying behavior on a large online dating site”, In Proceedings of 8th international AAAI conference on weblogs and social media, 2014.

Žakelj, T., “Internet dating and respectable women: Gender expectations in an untraditional partnership and marriage market-the case of Slovenia”, Sociologija, 56(1), 2014, pp. 5-21.

ARTICLE 15:

TWITTER AND THE POLITICAL LANDSCAPE – A GRAPH ANALYSIS OF GERMAN POLITICIANS

Authors: Annika Baumann, Humboldt-Universität zu Berlin, Berlin, Germany
Benjamin Fabian, Humboldt-Universität zu Berlin, Berlin, Germany
Stefan Lessmann, Humboldt-Universität zu Berlin, Berlin, Germany
Lars Holzberg, Humboldt-Universität zu Berlin, Berlin, Germany

Proceedings: 24th European Conference on Information Systems (ECIS'16)

Publication year: 2016

Abstract

This paper examines the Twitter social graph of German politicians and political parties during a time period not potentially biased by nearby elections. Based on a data set of 1,719 politicians across the entire political spectrum of this important country in the EU, two graphs are constructed, which also reflect relationships within and between parties: the follower graph, consisting of all follower-followee relationships between German politicians, and the "mention graph", which models direct references of politicians to their colleagues. Our main contributions are as follows: First, we analyse these graphs according to several statistics and graph metrics, characterizing political parties according to their collective participation in Twitter. We also investigate the openness for following ideas across political camps, resulting in the discovery of three distinct groups of political parties. We also find that membership in political parties itself explains only little of the variation in the formation of ties. There is also evidence that politicians with less activity exhibit a higher degree of openness than users with active engagement in tweets and discussions. This case study on social media adoption in politics leads to interesting insights into political debate in the information society.

***Keywords**: Twitter, Graph Analysis, Politics, Germany*

1. Introduction

Since its foundation in 2006, Twitter has become the most famous microblogging services worldwide with more than 300 million monthly active accounts in 2015 (Statista, 2015). The use and impact of social media among the political sphere is growing as well. Politicians use microblogging for direct interaction with colleagues and the electorate, comments on current political discussions and affairs as well as a medium for self-portrayal and marketing. There is a growing use of social media to foster the formation of public opinion and engaging in substantial political discussion (Tumasjan et al., 2010).

German political parties started to adopt online media on a larger scale just prior to the general elections in 2002 (Gibson et al., 2003). Nowadays, even though some German politicians have adopted microblogging via Twitter, others are still rather reluctant to use this new medium at a larger scope (Jungherr, 2013). This also causes a lack of research on the use of microblogging platforms among German politicians and their corresponding parties, since the focus of earlier studies was mostly on US and UK related campaigning activities on Twitter (Jungherr, 2014) despite Germany being one of the most influential countries in the EU in terms of politics and economics (WEF, 2015). Understanding the exchange of opinions among German politicians can therefore foster the study of global political developments.

Social media and especially microblogging platforms can be used to reach millions of potential voters and to directly interact with other politicians by creating an additional political platform for campaigns. In case of the US election in 2016 some of the presidential candidates had millions of followers in March 2016 (Statista, 2016), e.g., Donald Trump with 6.92 million followers and Hillary Clinton with 5.69 million followers. Since election candidates and political parties are in direct competition with each other, their interaction on microblogging platforms such as Twitter can attract immense attention. Thus, it is not only important to understand how politicians interact with externals via Twitter but also how they communicate and engage with each other on such an online platform.

Authors such as Gibson et al. (2003), Glassman et al. (2009) as well as Larsson and Moe (2012) ascertained that Internet activities facilitate the propagation of various political views when used for typical campaigning activities such as information dissemination and opinion formation. Furthermore, Utz (2009) finds that the usage was able to not only draw attention to their political program but also to increase the involvement of voters and the reputation of politicians. Therefore, reconstructing and analysing networks of politicians in online social media can lead to a deeper, empirical, and data-driven understanding of political exchange within and across political parties. Moreover, it could lead to a structural characterization of the internal and external exchange of particular camps. Not least, such analyses help to identify influential politicians and common themes that can form bridges between opposing parties, in particular if no elections are nearby that heat up the debate.

Furthermore, since the use of social media platforms by politicians was often in the focus of research during time frames of political booms such as election campaigns, we collected a unique dataset of a time span well-known for rather low frequent activities of political parties in Germany. This enables us to gain a deeper knowledge on how actively politicians behave on Twitter without the pressure of being overly present, i.e., their behaviour when there is no federal election nearby. This will be helpful to gain a deeper understanding on the pairwise relationships and interaction with their colleagues in times of political recessions. Therefore, our research contributes to understanding how politicians in Germany act in times when there is no event related to their activities on Twitter. A similar research has been conducted for the case of Australian politicians (Macnamara, 2011).

In order to gain a deeper and quantitative understanding on the interaction among politicians on Twitter and based on the suggestions of Yoon and Park (2014) as well as Hegelich and Shahrezaye (2014), we adopt recent methodological advances and conduct graph analyses by investigating two different graph structures – the *follower* graph, containing all connections to followers and followees, and the *mention* graph, consisting of all mentions of politicians by other German politicians. Mentions are part of tweets that contain the textual structure *@username*, which is explicitly referring to another user. These graphs are analysed with respect to reciprocity, clustering, density, homophily as well as further important statistical and structural measures. A particular focus will be placed on differences between German parties in their online activity and the willingness to engage in cross-party discussion.

The remainder of this paper is organized as follows: First, we will summarize related literature and give a brief introduction to the German political landscape. Afterwards, we will explain how we gathered the dataset that we used to generate the Twitter graphs. This is followed by the results of our graph analyses. We will conclude with a summary, limitations and an outlook on future work.

2. Related Work

Although the microblogging platform Twitter was founded not earlier than 2006, the use of Internet-based activities for political contexts has been known before. Following the US presidential elections in 2000 where the electoral campaign was supported by the use of different digital media channels, the first notable usage of Internet campaigning in Germany took place prior to the federal elections in 2002 (Gibson et al., 2003).

Mainly those related studies were relevant for our work that focused on politicians as entities of interest and used the microblogging platform Twitter as the main data source. Therefore, research on citizen engagement on Twitter in the political context was considered out of scope. As Table 1 shows, there is a lot of research on how politicians use Twitter focusing on either the US or the UK (Jungherr, 2014) – but looking at selected other countries on which the most research is available, evidence is clear that several studies analysed data prior to election campaigns. This is especially true for the case of Germany. In contrast, our research focuses on the analysis of German politicians' behaviour on Twitter in a non-event related time frame.

Furthermore, graph analysis is a method which so far has not been conducted that often (references marked with * in Table 1). In general, most of the studies using this method concentrate on the understanding of who interacts with whom (Jungherr, 2014). Other methods that have been applied are content analysis (e.g., Jackson and Lilleker, 2011) or descriptive analysis of tweets and hashtags, sentiment analysis (e.g., Plotkowiak and Stanoevska-Slabeva, 2013), experiments (e.g., Lee, 2013; Lee and Oh, 2013; Lee and Shin, 2012) or interview (e.g., Grussel and Nord, 2012). We will apply graph analysis methods to close this research gap and gain a better understanding of politicians on Twitter from a graph perspective.

In summary, most related research has studied countries such as the US and the UK without the methodological approach of graph analysis. Thus, we focus on these aspects and analyze follower-followee relationships as well as mentions among German politicians using graph analysis to close this research gap. Moreover, most of earlier research used time intervals prior to elections – especially in case of Germany, which could potentially bias the results. In contrast, our research aims to analyze political interaction on Twitter if there is no federal election nearby.

Country	Data Focus –Politicians, (Partial) Use of Twitter data	
	Campaign-related	**Not Campaign-related**
Australia	Bruns and Highfield, 2013* Macnamara, 2011	Grant et al., 2010* Macnamara, 2011
Asia		Hsu and Park, 2012 Kim and Park, 2011* Lee, 2013 Lee and Oh, 2012 Lee and Shin, 2012 Otterbacher et al., 2013 Yoon and Park, 2014*
Germany	Elter, 2013 Hegelich and Shahrezaye, 2014 Jungherr, 2010, 2012 Plotkowiak and Stanoevska-Slabeva, 2013* Plotkowiak et al., 2010 Thimm et al., 2012	Otterbacher et al., 2013 Thamm and Bleier, 2013
Netherlands	Broersma and Graham, 2012 Vergeer and Hermans, 2013 Vergeer et al., 2011, 2013	Verweij, 2012
Nordic Countries (DK, FI, NO, SE)	Grussel and Nord, 2012	Sæbø, 2011
United Kingdom	Broersma and Graham, 2012 Baxter and Marcella, 2013 Graham et al., 2013a, 2013b Jackson and Lilleker, 2011	Aharony, 2012 Otterbacher et al., 2013
United States of America	Ammann, 2011 Bode et al., 2011a, 2011b Christensen, 2013 Conway et al., 2013 Hanna et al., 2011 Mirer and Bode, 2013	Aharony, 2012 Chi and Yang, 2011 Glassmann et al., 2009 Golbeck et al., 2010 Hemphill et al., 2013 Hong, 2013 Lassen and Brown, 2011 Otterbacher et al., 2013 Peterson, 2012 Williams and Gulati, 2010

*Table 1. Overview of related work focusing on country of origin of the data (references marked with * used graph analysis methods in their work).*

3. Political Parties in Germany

The number of parties in the political system of Germany varies. The last federal election took place in 2013 with 30 different parties. Only a few of them reached the critical threshold of five percent to be represented in the German Bundestag. At the time of writing, the Bundestag consists of four different parties: The *Social Democratic Party*, the *Christian Democratic Union* in combination with the *Christian Social Union*, the *Left Party* and the *Green Party* (BPD, 2013). In order to depict a clearer picture of the political landscape, we will not only focus on those four parties but will also include the *Free Democratic Party*, the *Alternative for Germany* and the *Pirate Party*. All others will be aggregated into the pool *Other*.

The *Christian Democratic Union* (CDU) and the *Christian Social Union* (CSU) together constitute the conservative camp in the Germany. We will refer to these two parties by the abbreviation *CDU*. The core principles of the *CDU* are centered on traditional and Christian values of family, social cohesion, and harmony among different social classes (CDU, 2007). Some of their most important coalition partners are either the *Free Democratic Party* (FDP) or the *Social Democratic Party* (SPD).

The *Social Democratic Party* (SPD) is a traditional labour party addressing a large share of the German population. After the conservative group of *CDU* and *CSU*, it is the second largest party in Germany in terms of membership and political impact (Niedermayer, 2015). Its core values comprise justice and solidarity, aiming to establish equality in participation and opportunities (SPD, 2007). Their main coalition partners are the *CDU*, the *Left Party* and the *Green Party*.

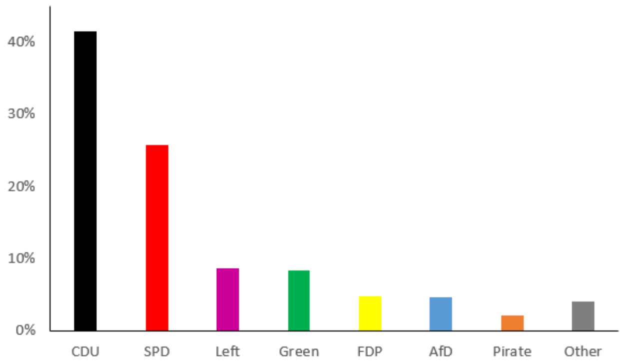

Figure 1. Election result of the last federal election in 2013.

The *Green Party* developed and established itself in the late 70s. Its main principles revolve traditionally around ecological and sustainable solutions in all aspects of life (Die Grünen, 2015).

The *Left Party* is rather socialistic and its core values revolve around solidarity. The party claims that participation and freedom can only be truly achieved when the economy is subordinated to solidarity and self-determination (Hildebrandt 2015). Their main coalition partners in the past have been either the *SPD* or the *Green Party*.

The *Free Democratic Party* (FDP) was not able to surpass the five percent clause in the federal election in 2013. Nonetheless, the *FDP* is represented in certain parliaments on the state level. The party's core values are mainly centered around personal freedom and personal responsibility (Patton, 2015). The party's former coalition partner was the *CDU*.

The *Pirate Party* was founded in 2006 and had initially a narrowed focus on Internet policy, focusing on freedom and data privacy. Their core principles are open access to education and public welfare systems as well as rejecting digital monitoring or data retention (Hebenstreit, 2015). Naturally, due to their digital focus, the *Pirate Party* members are often very active on Twitter.

The most recent party is *the Alternative for Germany* (AfD). Its core topics are constantly evolving from euro criticism to mainly national-conservative goals in recent times. The *AfD* is often criticized for being a right-wing populist party which enables it to attract mostly protest voters and conservatives but also strands of the political right (Lewandowsky, 2014).

4. Methodology

We decided to use a graph analysis approach to conduct our study of interactions of politicians on Twitter. A graph is a mathematical construct which consist of nodes and edges that connect two nodes. Nodes represent certain objects of interest and edges their interaction with each other. In our case, a node represents a specific politician. Depending on the selected graph, an edge reflects a different type of interaction. Based on the approach of Yoon and Park (2014) and suggested by Hegelich and Shahrezaye (2014), we decided to build two different graphs – the *follower* graph, containing all connections of followers and followees as well as the *mention* graph, which consists of all mentions of politicians by other German politicians. Both graphs are so-called directed graphs, which means that a specific interaction between two nodes is one-sided, e.g. two politicians do not have to necessarily follow each other since it is sufficient if only one follows the other.

We gathered data within a two-week period in August 2015, using the Twitter REST API. We decided to collect data originating from this time period since we wanted to understand how politicians use the communication platform Twitter during times where political actions are rather at a low frequency since the summer holiday takes place in Germany at this time span. The selection of this time frame helps us to understand if politicians are still actively using open communication tools to show presence to others such as potential voters as well as how they interact with each other.

Party	# Accounts	Average # / Median #		
		Tweets	Follower	Followees
AfD	26	465 / 50	186 / 63	1,007 / 106
CDU	421	1,441 / 285	274 / 106	1,428 / 218
Green	313	2,316 / 695	531 / 348	3,158 / 1,074
Left	171	2,093 / 391	418 / 200	2,728 / 543
FDP	165	1,580 / 482	522 / 213	1,294 / 443
Pirate	110	2,876 / 1,374	880 / 506	15,906 / 10,770
SPD	447	2,023 / 325	331 / 151	1,391 / 299
Other	66	814 / 197	239 / 105	2,482 / 449
Total	1,719	1,883 / 407	410 / 190	2,810 / 469

Table 2. Descriptive statistics of the retrieved dataset.

The selection of relevant politicians and their respective Twitter accounts was based on information taken from pluragraph.de (Pluragraph, 2015). This non-for-profit website depicts all non-commercial social media accounts of politicians, organizations as well as cultural and administrative entities. In August 2015, it listed 3,511 German politicians, of which 1,719 had an active Twitter account. Deleting all protected accounts, where access to tweets is only granted to Twitter users with permission, this resulted in 1,683 valid accounts in the final data set. The retrieved data contains information such as the number, content, and hashtags of tweets, the number of followers and friends as well as personal details, including name and Twitter membership data. To obtain a recent snapshot of the Twitter social graph, we downloaded the 200 most recent tweets of every considered politician, with hashtags and mentions of other politicians. Descriptive statistics of the dataset are shown in Table 2.

Based on the approach of Yoon and Park (2014), the analysis of the graphs considers only relations of nodes inside the data set to focus on the connections between German politicians. This means that only those nodes and edges are represented in the final graph which belong to the political sphere, i.e. nodes representing a German politician and edges containing a follower-relationship respectively mention between two politicians. Therefore all follower-relationships and mentions of a politician to another entity that is not a German politician, e.g., a news agency, a public person or a company, were deleted from the graph.

Furthermore, in the subsequent analysis, the mention graph will be compared to the follower graph. In order to facilitate this comparison, the follower graph should contain only the same source nodes as the mention graph, i.e., an interaction of both our defined graphs. Of the 1,719 politicians in the data set, 1,284 are found to be actively engaging in direct discussions with mentions, i.e., referring to other German politicians in their tweets. In summary, three different graphs are analysed:

- first, the full follower graph containing all follower-followee relationships between German politicians;
- second, the mention graph with all mentions of politicians regarding other colleagues;
- and third, the intersection graph, containing only the nodes that both exist in the follower as well as the mention graph.

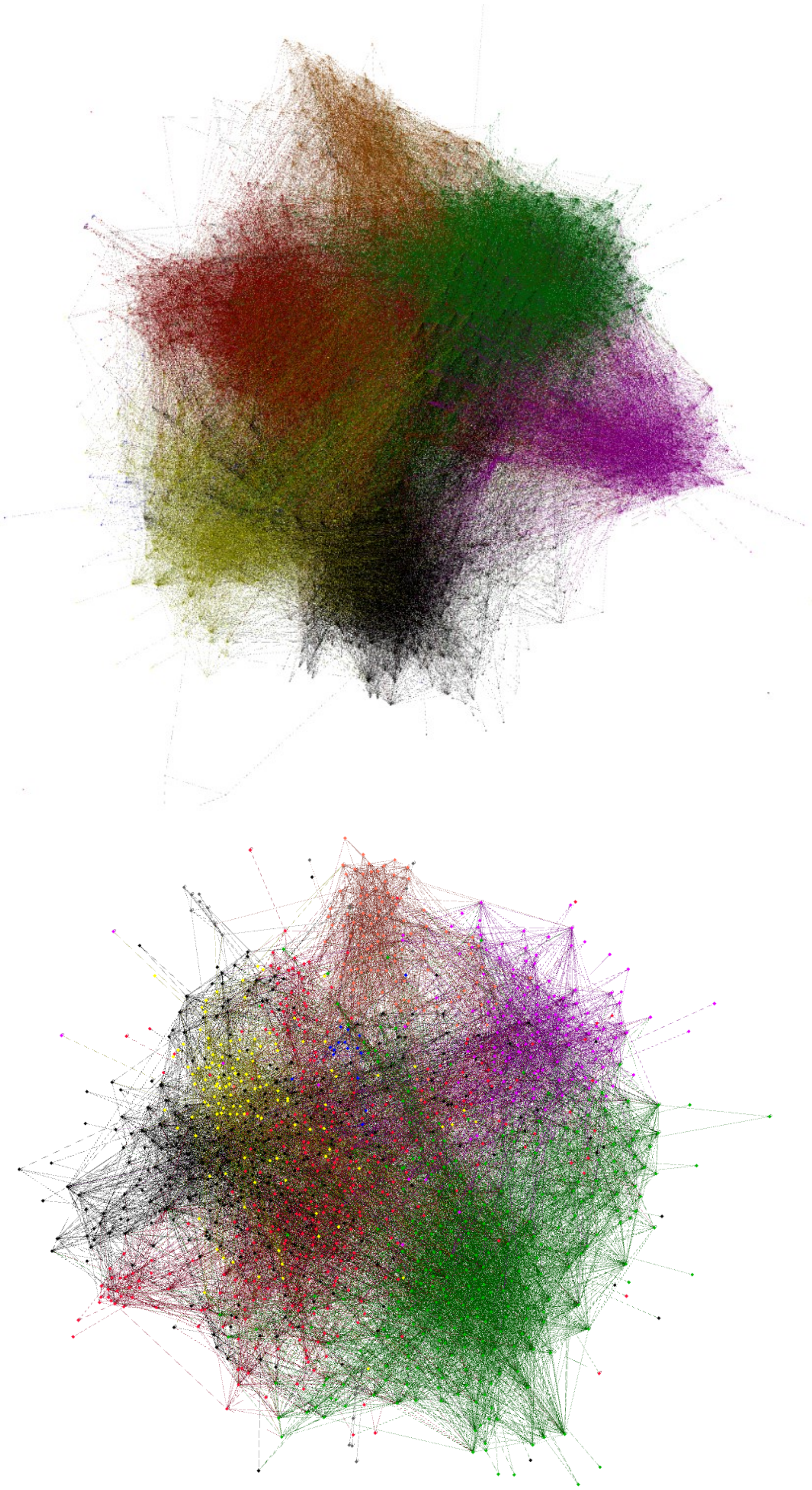

Figure 2. *Visualization of the follower (top) and the mention graph (bottom) [SPD: red, CDU: black, Green: green, Left: purple, FDP: yellow, Pirate: orange, AfD: blue].*

The final follower graph consists of 1,683 nodes and 99,792 edges. Furthermore, in total, 240,925 tweets have been retrieved. Of those, 51,375 tweets mention another German politician and 74% of politicians used the mentioning feature in their last 200 tweets. Finally, all self-references (789) as well as retweets (32,692) were deleted from the final data set, resulting in a mention graph consisting of 1,284 nodes and 17,501 edges.

	AfD	CDU	Green	Left	FDP	Pirate	SPD	Other
AfD	47	9	7	4	4	4	5	2
CDU	6	2,416	425	127	79	64	410	8
Green	9	384	4,522	206	50	118	566	7
Left	5	105	220	1,232	13	47	156	2
FDP	10	154	159	30	729	23	190	5
Pirate	1	72	105	40	31	1,023	122	23
SPD	6	338	442	138	61	82	2,891	7
Other	1	17	16	4	14	12	26	82

Table 3. Overview of inter-party follower relationships.

The graphical representation in Figure 2 of both graphs was depicted using the ForceAtlas2 algorithm with the help of the graph visualization software Gephi (Bastian et al., 2009).

5. Results

In the following, we provide analyses of the follower graph (FG), the mention graph (MG) and the intersection graph (FG-MG), which consists of FG reduced by the nodes not present in the MG. Visualizations of the FG and MG are shown in Figure 2. Nodes are allocated a specific colour depending on the party they belong to. The colour of an edge indicates an outgoing tie from a node of this particular party. Since the number of edges differs significantly (99,792 for FG, 17,501 for MG), the FG is a lot denser than the MG. It is visually apparent that parties are interwoven and interact with each other on Twitter. In case of MG, it seems that heterogeneity is more common.

Focusing on specific parties, the *CDU* has mostly internal connections and only a few ties with the *SPD* and *Green Party* (see also Table 3). The *Left Party*, in contrast, has a stronger tendency to also form connections with the *SPD* as well as the *Green Party*. The members of the latter two appear to be most active in both graphs in terms of internal and external ties. The *Green Party* rather dominates the FG, indicating that users of this group are most actively following users across all parties.

The MG depicts the *SPD* in the center, indicating that it receives and sends most mentions in tweets, both inside and outside of the party. Again, the *Green Party* appears to be most active but also with a focus on its own organization. The *FDP* has most of its ties with the *CDU* as well as the *SPD*. Furthermore, the *Pirate Party* is closely positioned to the *Green Party*. An intensive exchange of tweets and reciprocal relationships can be observed in this case.

5.1. Reciprocity and Clustering

Since the considered graphs are directed, we analysed whether ties between nodes are bi-directional or not, i.e., if users mutually follow one another. Reciprocity shows that in the majority of the cases the follower-relationship is only one-sided (see Table 4): For both, the FG and FG-MG, reciprocity is around 0.4, indicating that forty percent of the users in the graph are mutual followers. In case of the MG, only around 13 percent of all mentions are returned. This provides evidence that mentions are not used as bi-directional communication devices but to rhetorically accentuate messages.

The clustering coefficient is defined as the ratio of the number of edges existing between the neighbours of a node to the maximum possible number. The rather high number of 0.36 for the MG and 0.43 for the FG shows that several of the possible edges are realized and German politicians on Twitter are well connected with each other (see Table 4). To also take the size of existing cliques into account, i.e., sub-graphs with high density, we also computed the weighted clustering coefficient. This metric assigns higher weights to those sub-graphs that contain more nodes and edges. In case of this weighted clustering coefficient, the respective values are generally lower but the same trend remains. This gives evidence that indeed the political social graph on Twitter consist of several sub-graphs consisting of the specific political parties. In case of MG even if mentions are rather uni-directional as seen in the case of reciprocity, there are still several politicians who seem to be encouraged to mention others, maybe due to being triggered because another politician used this feature on themselves.

	Reciprocity	**Clustering Coefficient**	
		Overall	**Weighted**
FG	0.39	0.43	0.25
MG	0.13	0.36	0.18
FG-MG	0.40	0.41	0.27

Table 4. Reciprocity and clustering coefficient for each graph.

5.2. Centrality

An important measure to describe the structure of a graph is the degree, i.e., the number of nodes a node is connecting to. In our case this gives an indication of how often members of a specific party are being followed and how many other politicians they follow or respectively mentioned. Since all considered graphs are directed, they can be characterized by the in- and out-degree, referring to the number of followers and the number of followees or the number of ingoing and outgoing mentions, respectively. As shown in Table 5, except for the *AfD*, these values are similar for each party. However, the *Green Party* is better connected than all others. *Green Party* members are followed on average by 105.62 politicians and follow on average 105.53 other colleagues. Compared to the second largest political groups in the data set consisting of the *CDU* and the *SPD,* this results in 70 percent more ties on average for each *Green Party* member. Furthermore members belonging to the *Green Party* receive and give on average three more mentions compared to the next largest party.

	Out-Degree FG-MG	**In-Degree FG-MG**	**Out-Degree MG**	**In-Degree MG**
AfD	14.32	8.44	4.27	3.74
CDU	60.98	62.11	6.50	6.35
Green	105.53	105.62	9.62	9.72
Left	60.97	61.92	5.35	5.54
FDP	68.45	66.33	5.85	4.02
Pirate	62.87	58.33	7.62	7.96
SPD	66.35	68.03	5.82	6.60
Other	24.36	16.67	2.68	1.90

Table 5. Average degree centrality measures per party.

Since the degree only covers the direct neighbourhood of a node, we also considered the closeness centrality. This metric measures the average distance from a node to all other nodes, taking therefore the whole graph into account. The higher value, i.e., the closer it gets to one, the less separated a node is from others. As Table 6 shows, except for the *AfD*, the values are similar for in- and outward closeness across all parties. Nevertheless, the average member of the *Green Party* appears to be slightly more centralized in the political Twitter sphere.

In general, there seem to be three different clusters apparent based on centrality measures. The first one consists of the Green Party which seems to be best connected. The second cluster consists of the *AfD* and the party containing *Other* since compared to others they are not so well connected. All other parties not mentioned so far, i.e., *CDU*, *Left Party*, *FDP*, *Pirate Party* and *SPD*, are quite homogenous in terms of degree and closeness, having quite similar values and therefore a comparable behaviour on Twitter.

	Out-Close FG-MG	In-Close FG-MG	Out-Close MG	In-Close MG
AfD	0.38	0.35	0.16	0.12
CDU	0.42	0.42	0.15	0.16
Green	0.46	0.45	0.17	0.18
Left	0.42	0.41	0.15	0.16
FDP	0.43	0.42	0.15	0.15
Pirate	0.42	0.40	0.16	0.17
SPD	0.43	0.42	0.15	0.16
Other	0.36	0.36	0.14	0.10

Table 6. Average closeness centrality measures per party.

5.3. Density and Homophily

With respect to density, which gives the proportion of all theoretical connections to those that are actually present, we observe that for all graphs the overall density is rather low. The FG-MG has the largest value of 0.06. The density of the mention graph is lowest with 0.01, which is reasonable as it reflects politicians talking explicitly about other colleagues during the observation period.

Regarding the densities *within* parties focusing on follower-followee relationships, we find that the internal density in the *AfD* is by far the largest (0.17), followed by the *Green Party* with 0.11. All other parties range between 0.04 and 0.01. This gives evidence that both, *AfD* and the *Green Party*, have the strongest focus to connect with their own political sphere compared to all others to connect. Considering mentions, we observe a similar picture. Again the *AfD* ranks highest with a density value of 0.45. The second highest value is reached by the *Pirate Party* with 0.14. Again, all other parties have rather similar density values for mentions ranging between 0.08 and 0.02.

When calculating densities within and between groups for FG-MG, this observation becomes even more prevalent (see Table 7). Here we only consider those politicians who follow each other and mention others and the density is comparably larger than in FG. Again, the *AfD* has by far the largest density value of 0.45. Across the main diagonal the densities are consistently higher than elsewhere. This gives evidence that political parties are more likely to support their own party not only though following but also through mentions. Especially the AfD seems to have the habit to extensively support their inner circle since they are more prone to mention their colleagues.

	AfD	CDU	Green	Left	FDP	Pirate	SPD	Other
AfD	0.45	0.00	0.01	0.01	0.01	0.00	0.01	0.01
CDU	0.00	0.15	0.02	0.01	0.03	0.01	0.02	0.00
Green	0.00	0.02	0.28	0.03	0.02	0.03	0.03	0.01
Left	0.00	0.03	0.01	0.31	0.01	0.02	0.02	0.01
FDP	0.00	0.03	0.04	0.01	0.35	0.02	0.02	0.03
Pirate	0.00	0.03	0.01	0.03	0.03	0.36	0.02	0.03
SPD	0.00	0.03	0.02	0.02	0.02	0.01	0.14	0.01
Other	0.00	0.02	0.01	0.01	0.04	0.03	0.02	0.09

Table 7. Between and in-group densities of parties in FG-MG.

Using a bootstrap paired sample T-test with 5,000 samples, we tested for differences in the probabilities of a tie in FG-MG and MG. We have chosen to only test these two graphs since the nodes are the same and they are therefore easily comparable. The difference between densities of 0.05 (0.06 for FG-MG, 0.01 for MG) is found significant ($p < 0.01$, std < 0.01). We may therefore conclude that the densities of the FG-MG are significantly larger.

Since we found evidence that the densities within parties are relatively higher than between groups, we applied an Anova Density Test using a structural blockmodel to test for homophily (Hanneman and Riddle, 2015). This method verifies the differences between groups across a range of pre-defined clusters. In our case, these clusters are representing political parties. Importantly, the test does not impose explicit requirements on differences between clusters in order to verify them. We assume that this relaxed assumption better fits our dataset than other homophily blockmodels as there is no theoretical intuition to expect a recurring pattern across all groups.

We applied the test to FG as well as to MG with 5,000 permutations to gain robust results. Significant differences between parties for FG are enlisted in Table 8. However, although some deviations are most likely occurring not randomly the actual differences can be rather small. The calculated R-Square, indicating how homophily accounts for the variances in pair-wise ties, values to 0.01. Thus, despite some significant results, the overall degree of homophily in FG does not to account for the variance in pair-wise ties. Most notably, the differences in densities are most of the times significant for the *AfD* as well as the *Green Party*. Thus, political parties represent a certain form of clustering in the Twitter social graph, but they account only for a minor part of the observed variation in follower behaviour.

The homophily test of the mention graph finds significant results for all cases, proposing significant differences in mentioning behaviour across parties. However, the calculated R-Square of 0.01 indicates that the impact is rather minimal and other factors account for the observed variation.

The observation that FG-MG has a slightly larger density compared to FG in case of the whole graph as well as within party gives evidence that politicians who are being active in communication on Twitter are more connected overall, especially within their respective parties. Overall, the density of ties within the studied graphs is generally low but higher within individual parties.

	AfD	**Green**	**CDU**	**Left**	**FDP**	**Pirate**	**SPD**	**Other**
AfD	-	*	*	-	*	*	-	-
CDU	*	-	-	-	*	-	-	-
Green	*	*	*	*	*	*	*	-
Left	*	*	-	-	-	-	-	-
FDP	*	*	*	-	-	*	-	-
Pirate	*	*	*	-	*	-	-	-
SPD	-	*	-	-	-	-	-	-
Other	-	*	-	-	-	-	-	-

Table 8. Significant differences between party densities in FG (significant at .05%).*

5.4. Group-external and Group-internal Ties

To better understand if certain political parties have ties rather within their own boundaries or not, we calculated the *E-I index*, developed by Krackhardt and Stern (1988). This metric ranges from -1 to +1 mapping respectively to total closure or total openness. This can then be compared to a hypothesized value of a graph with a random distribution of ties having no propensity in either direction. The following calculations were based on a reasonably high number of 5,000 permutations.

Table 9 (left) gives an overview of the results on the level of the whole graph. FG-MG has a significantly different value than FG on which it is based, but is close to the value of MG. In case of FG, 77% of ties go outside the peer groups which is in line with the idea of users wanting to be informed by the whole political spectrum. At the same time, users actively mentioning others exhibit a propensity of keeping their ties closer in their group. This might also give evidence that mentions are rather used to support than to oppose other politicians.

Table 9 (right) considers the graph on party level. Notably, only the blocks of *Other* and the *AfD* show a positive sign for FG-MG and MG. Given the relatively small sample size of both groups, the results have to be interpreted carefully. For smaller parties, it is likely that the share of internal connections is relatively smaller having a less deteriorating effect on the E-I index.

	Intern	Extern	E-I	StD	E(E-I)
FG	0.23	0.77	0.54	0.01	0.62
MG	0.65	0.35	-0.29	0.01	0.63
FG-MG	0.64	0.36	-0.28	0.01	0.63

	FG	MG	FG-MG
AfD	0.91	0.30	0.10
CDU	0.49	-0.29	-0.22
Green	0.33	-0.41	-0.40
Left	0.83	-0.26	-0.21
FDP	0.79	-0.08	0.01
Pirate	0.86	-0.16	-0.43
SPD	0.53	-0.30	-0.24
Other	0.97	0.75	0.33

Table 9. *E-I index for whole graphs and for each political party.*

5.5. Krackhardt's Graph Theoretical Dimensions of Hierarchy

The previous analyses were based on the underlying idea of horizontal differentiation. However, it is also possible to study vertical differences in the form of hierarchies. Krackhardt et al. (1994) proposed the Graph Theoretical Dimensions of Hierarchy (GDT), providing four dimensions of hierarchy in directed graphs, which are interpretable as indices from 0 to 1 where a higher value points out a stronger presence of hierarchy. The baseline scenario assumes a purely hierarchical structure of one superior node having outward ties to all other nodes in the graph with an in-degree of zero. Krackhardt's indicators measure the deviation from this idealistic case.

The first variable, connectedness, indicates that at least one node is able to reach and connect to all other nodes, though not necessarily directly. The second item, hierarchy, shows whether hierarchies are revoked through reciprocal ties demonstrating equal status. Third, the efficiency variable denotes to which extent a graph exhibits no redundant or multiple paths but only has the least amount of ties to remain connected. The last item, least upper bound (LUB), indicates whether there are only few nodes dominating the majority of others. The resultant values are shown in Table 10.

	Connectedness	Hierarchy	Efficiency	LUB
FG	1	0.18	0.95	1
MG	1	0.5	0.99	0.95
FG-MG	1	0.05	0.92	1

Table 10. *Krackhardt GDT for each considered graph.*

The level of hierarchy is with 0.5 highest for the mention graph but ten times smaller for FG-MG. The first observation possibly reflects that tweets containing mentions are not returned automatically, resulting in nonreciprocal ties. A low hierarchy in the reduced follower network, on the other hand, is related to the idea that active users are relatively more likely to make followers to friends and vice versa, i.e., having bi-directional ties. The level of efficiency is relatively high for all studied graphs. LUB is close to 1 and notably high for all three graphs. This item combines the previous observations and states that most of the nodes are eventually linked to a single group of superiors. Krackhardt's GDT indicates therefore a decent level of hierarchy in the Twitter graph which is only partly counteracted by reciprocity.

6. Summary, Limitations, and Future Outlook

Based on the results of the various applied graph metrics applied to a not-campaign biased dataset we observe three different groups of parties. First, the *Green Party* appears to be most connected and active inside the political group of users. Although the *Pirate Party* surpasses the *Green Parties'* average number of tweets and followers, the case is different when restricting the group to political users, exclusively. Second, the *CDU*, *SPD*, the *Left Party* and *FDP* appear to be quite homogenous in terms of graph statistics. Given the relatively minor differences, it appears that these parties behave similarly in terms of networking and posting. The third group is consisting of the *AfD* and further parties, and indicates the lowest level of overall involvement in the Twitter network. On Twitter, the usage behaviour of the established parties is found to be rather homogenous. Thus, with the exceptions of the *AfD* and the *Green Party*, political users have a common approach on how to use Twitter in contact with peer politicians. Furthermore, we find a notable difference between the regular FG and the FG-MG, which excludes nodes not present in MG. Graph densities are higher in FG-MG, while at the same time the closeness of individual clusters increases. Hence, even though active Twitter users have a broader network in terms of mentioning other users in messages, their actual follower graph is closer, consisting of users with similar political beliefs. In contrast, inactive users appear to have a broader scope of followees and are not limited by their individual party, exhibiting a higher openness.

Interestingly, according to our various graph metrics applied, we find that party membership itself explains only little variation of the relationships in both graphs. Previous research from Plotkowiak et al. (2010) and Yoon and Park (2014) could only be partially confirmed. Although politicians tend to follow their own party, we only find three distinct groups for the eight studied parties. Given the diversity of political beliefs and organisation, this result is rather surprising. MG exhibits an increased amount of clustering. In line with Yoon and Park (2014), this might imply that politicians tend to support each other in active conversations. Our research is subject to some limitations. In general, the restrictiveness of the public Twitter API only allows the retrieval of a limited amount of Tweets. The data set was retrieved in a time span of only two weeks in August 2015. The dynamic evolution of the graph should therefore be investigated in future work. The list of Twitter accounts was based on data retrieved from a website, relying on the completeness of this information. There is also no guarantee that the actual person is really being in charge of a user account. Despite the fact that one third of accounts are verified on Twitter, there is always the possibility of misused accounts (Glassman et al., 2009).

In future work, an analysis could take all available factors into account, such as length of membership, gender or location. In addition, the studied user base contained also politicians who are currently not active in policy-making or in a parliament. A distinction of active and passive politicians could therefore be of interest. Moreover, so far we only considered tweets in the MG containing mentions of another German politician. This could be extended to international politicians. Furthermore, hashtags and contents of tweets have been out of the focus so far. In the next step, we would like to include this additional information. We also plan to additionally derive the sentiment of each Twitter post to observe the underlying motivation of each politician to maintain a Twitter account and mention others.

We also aim to study the *n-hop neighbourhood* of each politician, including the non-political sphere. Inspired by the study of Verweij (2012) who studied the Twitter relationships between politicians and journalists in the Netherlands, we would like to apply similar research for the case of German politicians. Importantly, the question whether lobbyism could be discernible on Twitter is of high political interest. A first look at the top 30 most followed non-political users in our underlying dataset reveals that this lists contains mostly newspapers and magazines, indicating a rather objective follower network. However, a future in-depth investigation could reveal different patterns.

In summary, our case study on microblogging adoption in German politics leads to several insights into the political debate in the information society and prepares the ground for analysing influencers and lobbyism in politics.

References

Aharony, N. (2012). “Twitter use by three political leaders: an exploratory analysis.” *Online Information Review* 36 (4), 587–603.

Ammann, S. L. (2011). Why Do They Tweet? The Use of Twitter by US Senate Candidates in 2010. SSRN Working Paper Series.

Bastian, M., Heymann, S. and M. Jacomy (2009). “Gephi: An open source software for exploring and manipulating networks”. ICWSM 8, 361-362.

Baxter, G. and R. Marcella (2013). “Do Online Election Campaigns Sway the Vote? A Study of the 2011 Scottish Parliamentary Election.” *Libri* 63 (3), 190–205.

Bode, L., Hanna, A., Sayre, B., Yang, J. H. and D. Shah (2011b). “Mapping the political Twitterverse: Finding connections between political elites.” In: *Political Networks Conference 2011*.

Bode, L., Lassen, D., Kim, Y. M., Shah, D., Fowler, E. F., Ridout, T. N. and M. Franz (2011a). “Social and broadcast media in 2010 Midterms: the expanding repertoire of Senate candidates’ campaign strategies.” In: *American Political Science Association 2011 Annual Meeting*.

Broersma, M. and T. Graham (2012). “Social media as beat: Tweets as a news source during the 2010 British and Dutch elections.” *Journalism Practice* 6 (3), 403–419.

Bruns, A. and T. Highfield (2013). “Political networks on Twitter: tweeting the Queensland state election.” *Information, Communication & Society* 16 (5), 667–691.

Bundeszentrale für Politische Bildung (BPD) (2013). URL. https://www.bpb.de/politik/grundfragen/parteien-in-deutschland/138671/anzahl-der-parteien (visited on 11/27/2015).

CDU (2007). *Freiheit und Sicherheit – Grundsätze für Deutschland; Party Programme*. URL: https://www.cdu.de/grundsatzprogramm (visited on 11/27/2015).

Chi, F. and N. Yang (2011). “Twitter adoption in Congress.” *Review of Network Economics* 10 (1), Article 3.

Christensen, C. (2013). “Wave-riding and hashtag-jumping: Twitter, minority ’third parties’ and the 2012 US elections.” *Information, Communication & Society* 16 (5), 646–666.

Conway, B. A., Kenski, K. and D. Wang (2013). “Twitter use by presidential primary candidates during the 2012 campaign.” *American Behavioral Scientist* 57 (11), 1596–1610.

Die Grünen (2015). Parteiprogramm. URL: https://www.gruene.at/partei/programm/parteiprogramm (visited on 11/27/2015).

Elter, A. (2013). “Interaktion und Dialog? Eine quantitative Inhaltsanalyse der Aktivitäten deutscher Parteien bei Twitter und Facebook während der Landtagswahlkämpfe 2011.” *Publizistik* 58 (2), 201–220.

Gibson , R. Römmele, A. and S. Ward (2003). “German parties and internet campaigning in the 2002 federal election”. *German Politics* 12 (1), 79-108, DOI: 10.1080/09644000412331307534.

Glassman, M., Straus, J. R. and C. J. Shogan (2009). “Social networking and constituent communication: Member use of Twitter during a two-week period in the 111th congress.” *Congressional Research Service* 7-5700.

Golbeck, J., Grimes, J. M. and A. Rogers (2010). “Twitter use by the U.S. Congress.” *Journal of the American Society for Information Science and Technology* 61 (8), 1612–1621.

Graham, T., Broersma, M. and K. Hazelhoff (2013a). Closing the gap: Twitter as an instrument for connected representation. In: The Media, Political Participation and Empowerment, ed. Richard Scullion, Roman Gerodimos, Daniel Jackson and Darren Lilleker. London, UK: Routledge pp. 71–88.

Graham, T., Broersma, M., Hazelhoff K. and G. van’t Haar (2013b). “Between broadcasting political messages and interacting with voters: The use of Twitter during the 2010 UK general election campaign.” *Information, Communication & Society* 16 (5), 692–716.

Grant, W. J., Moon, B. and J. B. Grant (2010). “Digital dialogue? Australian politicians’ use of the social network tool Twitter.” *Australian Journal of Political Science* 45 (4), 579–604.

Grussel, M. and L. Nord (2012). “Three attitudes to 140 characters: The use and views of Twitter in political party communications in Sweden.” *Public Communication Review* 2 (2), 48–61.

Hanna, A., Sayre, B., Bode, L., Yang, J. H. and D. Shah (2011). “Mapping the political Twitterverse: candidates and their followers in the Midterms”. In: *ICWSW 2011: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, ed. Nicolas Nicolov, James G. Shanahan, Lada Adamic, Ricardo Baeza-Yates and Scott Counts. Palo Alto, CA: Association for the Advancement of Artificial Intelligence (AAAI), pp. 510–513.

Hanneman, R. A. and M. Riddle (2005). Introduction to social network methods. University of California Riverside.

Hebenstreit, J. (2015). *Piratenpartei Deutschland*. URL: http://www.bpb.de/politik/grundfragen/parteien-in-deutschland/42193/piraten (visited on 11/27/2015).

Hegelich, S. and M. Shahrezaye (2014). “The communication behavior of German MPs on Twitter: Reaching to the converted and attacking enemies.” In: *MPSA Conference*.

Hemphill, L., Otterbacher, J. and M. A. Shapiro (2013). “What’s Congress Doing on Twitter?”. In: CSCW 2013: *Proceedings of the 2013 Conference on Computer Supported Cooperative Work Companion*, ed. Amy Bruckman, Scott Counts, Cliff Lampe and Loren Terveen. New York, NY: ACM, pp. 877–886.
Hong, S. (2013). “Who benefits from Twitter? Social media and political competition in the U.S. House of Representatives.” *Government Information Quarterly* 30 (4), 464–472.
Hsu, C. -L. and H. W. Park (2012). “Mapping online social networks of Korean politicians.” *Government Information Quarterly* 29 (2), 169-181.
Jackson, N. A. and D. G. Lilleker (2011). “Microblogging, constituency service and impression management: UK MPs and the use of Twitter.” *The Journal of Legislative Studies* 17 (1), 86–105.
Jungherr, A. (2010). “Twitter in politics: lessons learned during the German Superwahljahr 2009”. In: *Workshop on Microblogging at CHI10* (ACM Conference on Human Factors in Computing Systems).
Jungherr, A. (2012). “Online campaigning in Germany: the CDU online campaign for the General Election 2009 in Germany.” *German Politics* 21 (3), 317–340.
Jungherr, A. (2013). “Schleppender Beginn: Deutsche Politiker entdecken Twitter nur zögerlich.” *Internationale Politik*, 54–59.
Jungherr, A. (2014). “Twitter in politics: a comprehensive literature review”. Available at SSRN 2402443.
Kim, M. and H. W. Park (2011). “Measuring Twitter-based political participation and deliberation in the south Korean context by using social network and triple helix indicators.” *Scientometrics* 90 (1), 121-140.
Krackhardt, D. and R. N. Stern (1988). “Informal networks and organizational crises: An experimental simulation.” *Social Psychology Quarterly* 123-140.
Krackhardt, D., Carley, K. M. and M. J. Prietula (1994). “Graph theoretical dimensions of informal organizations”. *Computational Organization Theory* 89 (112), 123-140.
Larsson, A. O. and H. Moe (2012). “Studying political microblogging: Twitter users in the 2010 Swedish election campaign.” *New Media & Society* 14 (5), 729-747.
Lassen, D. S. and A. R. Brown (2011). “Twitter: the electoral connection?” *Social Science Computer Review* 29 (4), 419–436.
Lee, E.-J. (2013). “Effectiveness of Politicians’ Soft Campaign on Twitter Versus TV: Cognitive and Experiential Routes.” *Journal of Communication* 63 (5), 953–974.
Lee, E.-J. and S. Y. Oh (2012). “To personalize or depersonalize? When and how politicians’ personalized tweets affect the public’s reactions.” *Journal of Communication* 62 (6), 932–949.
Lee, E.-J. and S. Y. Shin (2012). “When the medium is the message: How transportability moderates the effects of politicians’ Twitter communication.” *Communication Research* 41 (8), 1088-1110.
Lewandowsky, M. (2014). “Alternative für Deutschland (AfD): A new actor in the German party system.” *Friedrich Ebert Stiftung*.
Macnamara, J. (2011). “Pre and post-election 2010 online: What happened to the conversation?” *Communication, Politics & Culture* 44 (2), 18–36.
Mirer, M. L. and L. Bode (2013). “Tweeting in defeat: How candidates concede and claim victory in 140 characters.” *New Media & Society* 17 (3), 453-469.
Niedermayer, O. (2015). „Parteimitglieder in Deutschland: Version 2015.“ *Arbeitshefte Aus Dem Otto-Stammer-Zentrum* 25.
Otterbacher, J., Shapiro, M. A. and L. Hemphill (2013). “Interacting or just acting? A case study of European, Korean, and American politicians’ interactions with the public on Twitter.” *Journal of Contemporary Eastern Asia* 12 (1), 5–20.
Patton, D. F. (2015). “The prospects of the FDP in comparative perspective: Rest in peace or Totgesagte leben länger?” *German Politics* 24 (2), 179-194.
Peterson, R. D. (2012). “To tweet or not to tweet: Exploring the determinants of early adoption of Twitter by House members in the 111th Congress.” *The Social Science Journal* 49 (4), 430–338.
Plotkowiak, T. and K. Stanoevska-Slabeva (2013). “German politicians and their Twitter networks in the Bundestag election 2009.” *First Monday* 18 (5).
Plotkowiak, T., Ebermann, J. and K. Stanoevska-Slabeva (2010). “A longitudinal social network analysis of German politicians' Twitter accounts.” In: *Sunbelt Annual Conference*, Riva del Garda.
Pluragraph (2015). Pluragraph – Social-Media-Benchmarking and Social-Media-Analyse. URL: https://pluragraph.de/ (visited on 11/27/2015).
Sæbø, Ø. (2011). “Understanding TwitterTM Use among Parliament Representatives: A Genre Analysis”. In: *Electronic Participation: Proceedings of the Third IFIP WG 8.5 International Conference, ePart 2011*, ed. Efthimios Tambouris, Ann Macintosh and Hans de Bruijn. Heidelberg, DE: Springer, pp. 1–12.
SPD (2007). *Grundsatzprogramm*. URL: http://www.spd.de/partei/grundsatzprogramm/ (visited on 11/27/2015).
Statista (2015). “Number of monthly active Twitter users worldwide from 1st quarter 2010 to 3rd quarter 2015.” URL: http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/ (visited on 11/27/2015).

Statista (2016). “Number of Twitter followers of 2016 U.S. presidential candidates, as of March 15, 2016.” URL: http://www.statista.com/statistics/509579/twitter-followers-of-2016-us-presidential-candidates/ (visited on 04/05/2015).
Thamm, M. and A. Bleier (2013). “When politicians tweet: A study on the members of the German Federal Diet”. arXiv preprint arXiv:1305.1734.
Thimm, C., Einspänner, J. and M. Dang-Anh (2012). “Twitter als Wahlkampfmedium: Modellierung und Analyse politischer Social-Media Nutzung.” *Publizistik* 57 (3), 293–313.
Tumasjan, A., Sprenger, T. O., Sandner, P. G. and I. M. Welpe (2010). “Predicting elections with Twitter: What 140 characters reveal about political sentiment.” In: *International AAAI Conference on Web and Social Media (ICWSM)* 10, 178-185.
Utz, S. (2009). “The (potential) benefits of campaigning via social network sites.” *Journal of Computer-Mediated Communication* 14 (2), 221-243.
Vergeer, M. and L. Hermans (2013). “Campaigning on Twitter: micro-blogging and online social networking as campaign tools in the General Elections 2010 in the Netherlands.” *Journal of Computer-Mediated Communication* 18 (4), 399–419.
Vergeer, M., Hermans, L. and S. Sams (2011). “Is the voter only a tweet away? Micro–blogging during the 2009 European Parliament election campaign in the Netherlands.” *First Monday* 16 (8).
Vergeer, M., Hermans, L. and S. Sams (2013). “Online social networks and micro-blogging in political campaigning: the exploration of a new campaign tool and a new campaign style.” *Party Politics* 19 (3), 477–501.
Verweij, P. (2012). “Twitter links between politicians and journalists.” *Journalism Practice* 6 (5-6), 680-691.
WEF (2015). *World Economic Forum’s Global Competitiveness Report 2014-2015*. URL: http://www.weforum.org/content/top-10-most-competitive-economies-europe-2 (visited on 11/27/2015).
Williams, C. B. and G. J. Gulati (2010). “Communicating with constituents in 140 Characters or less: Twitter & the diffusion of technology innovation in the United States congress.” In: *Midwest Political Science Association Annual Conference 2010*
Yoon, H. Y. and H. W. Park (2014). “Strategies affecting Twitter-based networking pattern of South Korean politicians: Social network analysis and exponential random graph model.” *Quality & Quantity* 48 (1), 409-423.

ARTICLE 16:

PRIVACY ON REDDIT? TOWARDS LARGE-SCALE USER CLASSIFICATION

Authors: Benjamin Fabian, Humboldt-Universität zu Berlin, Berlin, Germany
Annika Baumann, Humboldt-Universität zu Berlin, Berlin, Germany
Marian Keil, Humboldt-Universität zu Berlin, Berlin, Germany

Proceedings: 23rd European Conference on Information Systems (ECIS'15)

Publication year: 2015

Abstract

Reddit is a social news website that aims to provide user privacy by encouraging them to use pseudonyms and refraining from any kind of personal data collection. However, users are often not aware of possibilities to indirectly gather a lot of information about them by analyzing their contributions and behaviour on this site. In order to investigate the feasibility of large-scale user classification with respect to the attributes social gender and citizenship this article provides and evaluates several data mining techniques. First, a large text corpus is collected from Reddit and annotations are derived using lexical rules. Then, a discriminative approach on classification using support vector machines is undertaken and extended by using topics generated by a latent Dirichlet allocation as features. Based on supervised latent Dirichlet allocation, a new generative model is drafted and implemented that captures Reddit's specific structure of organizing information exchange. Finally, the presented techniques for user classification are evaluated and compared in terms of classification performance as well as time efficiency. Our results indicate that large-scale user classification on Reddit is feasible, which may raise privacy concerns among its community.

Keywords: *Privacy, Social Media, Reddit, User Classification, Machine Learning*

1. Introduction

The popularity of social media and social networking has risen continuously since its first appearance. Users generate massive amounts of content on Facebook, Twitter, and similar social networking sites as well as on blogs, video sharing platforms, etc. Many companies are searching for opportunities in analyzing social media entries (Kaplan and Haenlein, 2010). This ranges from automatic processing of product reviews and opinions especially with the social components of online retailers (Oelke et al., 2009; Popescu and Etzioni, 2007) to predictions of the financial market (Ferguson et al., 2009) and even mining companies' market structures (Netzer et al., 2012). Also social sciences increasingly use the information contained in social websites. Especially political movements such as the Arab Spring (Lotan et al., 2011) or Occupy Wall Street (Tremayne, 2014) have been studied with the help of content analysis of social media.

Many of these analyses require knowledge of certain demographics such as the origin of the authors of messages. However, many users might begin to share less information about their profiles as anonymity and privacy are gaining more attention during the last years. In particular, the so called NSA leak in which highly confidential documents about online user surveillance by the US government were unveiled by Edward Snowden in 2013, increased the awareness that any information shared on the Internet could be abused by third parties. Therefore, many users search for opportunities to express themselves on social platforms without revealing their identity.

Reddit[20] is a bulletin board system where users can share news and posts from funny to controversial (Koh, 2013). When founded in 2005 it only had content in form of links or text entries and an up-and-down voting system for submissions. Soon a comment function was added that increased social interaction of users. Reddit pays a lot of attention to users' privacy, granting them the choice of disclosing none of their information except for the posts and comments they provide. This complicates large-scale content analyses compared to other social media platforms. So far, no research has investigated how well user privacy on Reddit is protected against sophisticated approaches to identify latent personal attributes.

In order to address this research gap, this article is dedicated to classifying users on Reddit based on their comments regarding their social gender as well as the more comprehensive classification into citizenship. Because of its supposed anonymity, using Reddit for content analysis is not yet widely applied. Two potential examples of usage possibility will be given in the following to motivate user classification in Reddit. First, Reddit contains many discussions on consumer products that can provide insights on the various segments of customers. Demographic information supports the customer analysis. Second, discussions on political issues are also quite frequent. Information on the origin of users might help to understand and follow the discourses and opinions presented. In particular, we focus on answering the question whether it is possible to classify users on Reddit simply based on the post and threads they have commented on.

The remainder of our article is structured as follows. First, we present a review of related work on classifying users and extracting latent information in social media. Then, our research method and a newly extracted text corpus are presented. Afterwards, a selection of state of the art classifiers is introduced and their performance is evaluated. Finally, we discuss limitations and present conclusions.

[20] URL: http://www.reddit.com

2. Related Work

User classification, and in particular identifying differences between genders in online and offline communication, has been a well-studied subject (Baumann et al., 2015). Already in the early 1970's, researchers made increasing use of statistics on speech patterns to find differences between the two genders (Lakoff, 1972; Trudgill, 1972) and to analyze women's position in society. With the emergence of electronic communication such as email and Internet chat, research also shifted focus to these areas. Herring (1996) analyzed gender differences in terms of emails and later on also in Internet chat conversations (Herring, 2000; Panyametheekul and Herring, 2003).

Schler et al. (2006) presented findings on gender and age differences between bloggers and developed a classifier which could predict gender at 80.1% and age of three classes at 43.8% accuracy. They used the (multi-class) Real Winnow classification algorithm on features of blogs such as word frequencies, categories of words as well as writing style. Their corpus contained information of 37,000 bloggers. Gender classification on bloggers was also performed by Yan and Yan (2006) using a naive Bayes classifier. They relied on features of a bag-of-word approach based on the content and additionally included the choice of font, punctuation marks as well as emoticons. For their corpus of 3,000 bloggers, they reported an F-measure of 68% and showed that it increases monotonically with the corpus size. Mukherjee and Liu (2010) classified the gender of bloggers achieving a decent accuracy of 88.56% by using *support vector machines* (SVM) regression on a well-chosen set of features. They selected and graded them using an ensemble feature selection algorithm that ranks the shares of all features on the overall performance by applying measures such as cross-entropy or mutual information.

Rao et al. (2010) build a scientific Twitter corpus to execute user classification. The authors used binary SVM for classification of gender, age, regional origin, and political affinity. They applied it to three kinds of features: socio-linguistic characteristics, unigrams and bigrams of tweets and a combination of these. For gender, the classifier showed a prediction accuracy of 72.3%, for age of 74.1%, for regional origin of 77.1% and for political orientation of 82.8%. Pennacchiotti and Popescu (2011a,b) also focused on user classification on Twitter. They looked at political orientation, affinity for a specific business (Starbuck's), and African-American ethnicity. They applied Gradient Boost Decision Trees to features such as profile information, sentiment of tweets and user behavior. Furthermore, they used topics of latent Dirichlet allocation (LDA) as features, applying two different versions: One was trained on the set of all users, the second one was only trained on a domain-specific training set (e.g., only users with labels of political affinity when classifying Democrats or Republicans). Their results for the classification case of ethnicity are rather imprecise at an F-measure at 65.5% whereas Democrats could be predicted at 91.5% and Republicans at 84.0%. Fans of Starbuck's were identified with an F-measure of 76.1%. Burger et al. (2011) managed to discriminate gender on Twitter at a comparably high accuracy of 92.0%. They chose Winnow2 as their classification algorithm. Their feature set consisted of word and character *ngrams* of the tweet contents, user profile information, user screen names, and user full names. Compared to a manual classification task using the Amazon Mechanical Turk, the automatic classification performed with a higher accuracy.

Facebook has also been the interest of gender discrimination in research as Wang et al. (2013) looked at one million random English status updates and applied LDA topic modeling, while identifying 25 topics which they labeled. With separation of users over 25 and under 25 years they could show differences in occurrence frequency of each topic for the different genders.

3. Research Method

Though English is the dominant language on Reddit, it would be useful to classify users independent of the language they are writing in. This certainly excludes the application of full text analysis to extract features for classification. But as a social network, Reddit also offers non-textual elements that can be exploited, in particular the structure of Subreddits and Reddits together with the comments and (dis)likes of users. Our main research question is therefore the following: Is it possible to automatically extract latent user attributes such as gender and citizenship from Reddit by classifying users based on their comments? Our corresponding method essentially consists of five process steps that are illustrated in Figure 1 which will be explained in more detail in the following.

Figure 1. Process of user classification on Reddit.

To gather a text corpus, we extracted all comments for several users. The selected usernames were provided by a publicly available list which was constructed by collecting 660,464 comments of users in November 2013 posted within nine days (Redditor twentythree-nineteen, 2013). Given this user list, the complete set of comments and votes for each user was then collected using the Reddit API (Reddit Inc., 2014). By this procedure we retrieved comments for 76,767 users. It has to be noted that with this approach the user names were not drawn independently from the list of all registered users on Reddit because there is no method to access random users without knowing their names. For further processing, the corpus was preprocessed to ensure good data quality. For this, we excluded all users without any comments. Although the corpus has originally been sampled from names of users who have commented, users are able to delete their comments and votes.

The next step consisted of pre-tagging the gathered corpus. Supervised learning requires a training set in order to make predictions. The input data for training needs to consist of the prediction's input (the information the prediction is based on) as well as the outcome (target) that the prediction should generate based on the input data. The outcome or label for each data entry is the user's classified attribute. For the purpose of this article, social gender with the label classes "female" and "male" and citizenship which is grouped by continent into "Africa", "Asia", "Australia and Pacific", "Europe", "North America", and "South and Middle America" are the labels of choice.

Since the Reddit corpus does not provide these annotations and we could not identify any available Reddit corpus that provides additional user information, we derive and use probable labels. They might be not a 100 % correct representation of real Reddit user attributes. To derive more accurate labels, manual tagging should be applied which can be based on asking users explicitly as well as looking individually at each of their comments. For the automatic derivation of labels, no machine learning should be used in order to preserve independence of labeling and prediction. Pennacchiotti and Popescu (2011a) introduced an algorithm for Twitter user classification based on regular expressions. They proposed the following pattern as a simple method to extract users' age and ethnicity from the user's comments: "*(I|i)(m|am|´m)[0-9]+(yo|year old) white(man|woman|boy|girl)*" (Pennachiotti and Popescu, 2011a, p. 282).

For our purpose of extracting users' citizenships, the labeling is based on regular expressions and extracts country names and demonyms using the "List of adjectival and demonymic forms for countries and nations" of Wikipedia (Wikipedia, 2014). For gender, a similar approach is applied suggested by Rao et al. (2010) which focuses on finding the possessive "my" in comments together with a clearly identifying terms, e.g., "husband", "girlfriend", or "hubby".

Since quotations can easily distort the label procedure, we use a feature of the Reddit API which provides a recommended style for quoting inside comments and enables easy removal of the quotations. It might be the case that users that do not adhere to that rule may be potentially misclassified. In order to reduce the number of false positives (for example, a woman being classified as "male"), all ambiguous users have not been included in the pre-tagged corpus (8 for citizenship; 122 for gender).

The final numbers of users for each label can be found in Table 1. A survey conducted on Reddit users (Reddit Inc., 2011) suggests similar relative numbers for gender (female: 18.8%, male: 81.1%) although a more recent survey based on Reddit's Internet traffic does not confirm these numbers instead suggesting a less strong difference (female: 36.3%, male: 63.7%). Either the traffic is not indicative for active users or Reddit became more popular among women within the time between the survey and the Internet traffic acquisition.

Gender	Users	Share
Male	19,991	78.5 %
Female	5,474	21.5 %
Total	25,465	100.0 %

Citizenship	Users	Share
Europe	5,613	37.4 %
North America	5,584	37.2 %
Asia	1,669	11.1 %
Australia and Pacific	1,173	7.8 %
South and Middle America	713	4.8 %
Africa	260	1.7 %
Total	15,012	100.0 %

Table 1. Number and relative share of tagged users by gender and citizenship.

A comparison to the 2011 survey (Reddit Inc., 2011) does not reveal similar relative numbers in terms of citizenship. Instead, North America is with 73.6% the strongest continent and Europe is with 17.2% only on the second position far behind (Pacific: 4.6%, Asia: 2.5%, South America: 1.6%, Africa: 0.4%). Recent reports provided by Alexa.com support the dominance of North American users above two-thirds of overall site visitors (Alexa Internet, 2014). One reason for the difference could be that particularly non-North American users will state their origin to set themselves apart since Reddit is a North American dominated website. Therefore, the labeled corpus of citizenship cannot be regarded as representative as a random sample.

Once the users are labeled in terms of gender and citizenship concerning the test set, a bag-of-words is created which is a representation of a document that disregards sequential order of its words. It can be considered as a table of words and the count each word appears inside this document. For our classification, users are treated as documents and Reddits or Subreddits as words. Therefore, the bag-of-(Sub)Reddits considers the number of all (Sub)Reddits a user commented on.

The actual user classification is the final step. It is conducted by using classification algorithms and evaluating their performance on the given test corpus. All the previous steps are providing the necessary input of the bag-of-(Sub)Reddits and the users' attribute labels. The user classification only considers one attribute at a time and does not include any statistical dependencies between the two attributes of gender and citizenship.

4. Classifiers

In our study, two families of state-of-the art classifiers were adopted: weighted soft-margin SVM classifiers and supervised LDA. SVM are a group of regression and classification algorithms that base on the idea of introducing a margin around the decision boundary to achieve a higher amount of generalization (Bishop et al., 2006, p.362). The support vector classification is a linear binary classifier and can therefore only discriminate between two classes. Its classification boundary is a separating hyper plane. A soft-margin permits a classification error in the training data to a certain degree and usually facilitates a better generalization (Bishop et al., 2006, p.331). When looking at the data of our corpus (Table 1), both attributes show strong imbalances among the classes. To tackle the disproportion, a weighted SVM was introduced (Osuna et al., 1997).

The second classifier family is related to generative models, which are statistical models that try to capture an underlying causal process (Bishop et al., 2006, p.366). This allows a modeling of real world problems and derivation of statistical solutions by inferring latent information. For the case of user classification, it is possible to model the process of generation of comments by users. For example, a user chooses a topic to write about and based on this she selects a (Sub)Reddit to comment. Thereby, the topic is bound to the author's personality or attributes such as gender or origin that influences the choice of (Sub)Reddits. Therefore, we are trying to utilize this causality to classify a user.

The generation of topics to exploit hidden information of a document was first applied by Furnas et al. (1988) by introducing latent semantic indexing which is in machine learning usually called latent semantic analysis (LSA). Its algorithm is relying on the usage of singular value decomposition of the term-document matrix and reducing it to a term-topic matrix and a document-topic matrix. Each document therefore has a certain "amount" (including zero) of every topic which describes its intrinsic latent structure. This idea was later extended to a statistically grounded version: probabilistic latent semantic analysis (pLSA). Latent Dirichlet allocation (LDA) is an extension of pLSA that basically adds a Dirichlet prior to the topic choice and the topics' word emissions. As LDA is an unsupervised learning algorithm, the generated topics could be orthogonal to the user attributes that should be identified (e.g., topics that distinguish age groups while the classifier would discriminate gender). It would be possible to use the topics generated by LDA as features and learn them using, for example, an SVM classifier (Pennacchiotti and Popescu, 2011a,b).

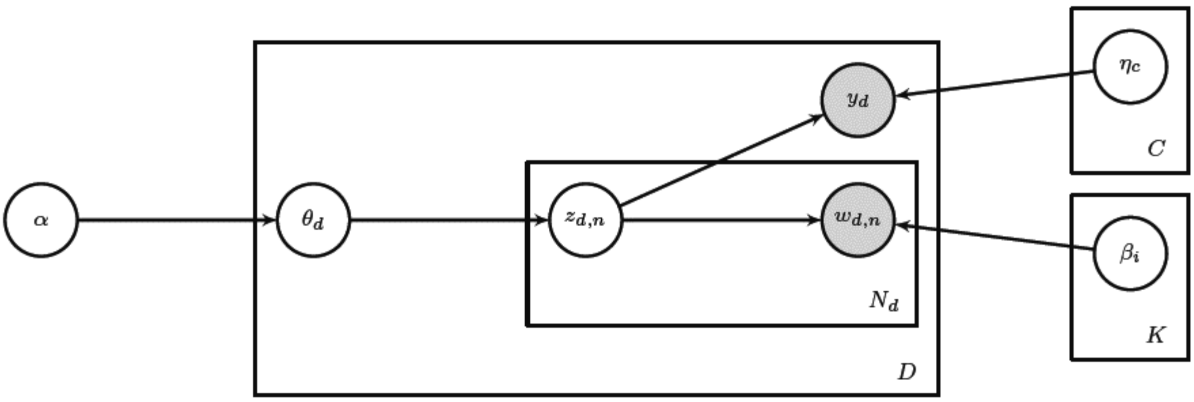

Figure 2. Plate notation of supervised latent Dirichlet allocation based on McAuliffe and Blei (2007). The outer plate (rectangular box) represents documents. The inner plate shows the choice of topics and words within a document.

But due to LDA's unsupervised nature it withholds the possibility to control the topic creation and therefore the ability to generate topics that – when learned – minimize the prediction error. Therefore it is useful to have an algorithm that supports supervision. We adopt a model called *supervised latent*

Dirichlet allocation (sLDA) (McAuliffe and Blei, 2007; Figure 2) which works as follows: Each topic is "emitting" certain terms as a multinomial distribution. Its prior is the Dirichlet-distributed random variable β_i which defines the probabilities for each term for each topic, and whose values are affected by its prior/parameter α_β. For each document, a mixture of topics is selected by sampling θ as a symmetric Dirichlet distribution parameterized by α_θ. The topic mix is then used to choose one topic for each word in the current document by sampling $z_{d,n}$ from the multinomial distribution with prior θ. Based on the chosen topic, a term inside the vocabulary is selected by the multinomial distribution $w_{d,n}$. The sLDA's model is very similar to that of LDA; only the additional random variables *Y* for each document with an additional parameter η for each possible label class are added. The distinct number of classes is *C* as used in the generative process for sLDA (McAuliffe and Blei, 2007):

1. Draw word probabilities (for each topic *i*): $\beta_i|\alpha_\beta \sim Dir(\alpha_\beta)$
2. For each document d in [1,D]:
   a. Draw topic proportion $\theta_d|\alpha_\theta \sim Dir(\alpha_\theta)$.
   b. For each term *n* in [1, N_d]:
      i. Draw topic assignment: $z_{d,n}|\theta_d \sim Mult(\theta_d)$. Actually, this is a categorical random variable. But for simplicity all categorical distributions will be regarded as one-trial multinomial random variables. This has the advantage that vectors of the form (0, 0, ..., 1, ..., 0, 0), with one dimension set to 1 and the others to 0, can be used instead of a scalar value.
      ii. Draw word: $w_{d,n}|z_{d,n}, \beta \sim Mult(\beta_{z_{d,n}})$.
   c. Draw response variable $y_d|z_{d,1:N}, \eta \sim Mult(\eta^T \overline{z_d})$ where $\overline{z_d} = \frac{1}{N}\sum_{n=1}^{N} z_n$ . .

In the original paper (McAuliffe and Blei, 2007) the response type is only restricted to a generalized linear model (GLM) and they based specific implementation on a normal distribution. But for the Reddit user classification a discrete response type is more appropriate. Therefore we focus on a multinomial variable y_d that regresses on $\overline{z_d}$.

5. Results

As performance indicators for classification of gender, *receiver operating characteristic* (ROC) and *area under the curve* (AUC) are chosen since they are advantageous over F-measure and accuracy for imbalanced data classes (Fawcett, 2006). For ROC analysis, a graph is plotted as point pairs of the false positive (FP) rate and true positive (TP) rate over a varying cut-off threshold of class probability for each data point of the test dataset. Thus, the threshold walks from zero to one and, depending on the probability of each test data point belonging to a respective class, the FP rate and TP rates change. A FP in case of the gender attribute might be a male user who was falsely classified as female or the other way around, whereas a TP indicates a correctly classified user in terms of gender.

AUC is a good measure to summarize an ROC curve by a single number, the definite integral from zero to one. A perfect classifier has an AUC value of 1.0 (or 100%), a random algorithm would score with an AUC of approximately 0.5 (or 50%). Therefore the larger the area under the ROC curve is, the better the classifier performs in the sense that an increase in the number of TPs does not lead to a likewise increase of the FP rate.

ROC analysis has the drawback that it only regards two-class classification problems. But for citizenship, six classes need to be considered. One solution would be to create several ROC plots for each class against all other classes (Fawcett, 2006). But due to its high-dimensional surface, it is visually intractable. Therefore we will not consider multi-class ROC but will instead focus on multi-class AUC.

Hand and Till (2001) suggested a multi-class AUC that is based on two-class AUC over all possible pairs of classes including their counterparts (class pair *(X, Y)* as well as *(Y, X)*). They calculate the overall AUC as:

$$A_{total} = \frac{1}{C(C-1)} \sum_{i \neq j} A_{i,j} \qquad (1)$$

Accordingly, performance of citizenship classification will be only depicted using AUC.

The following results were all acquired using a 6-fold cross-validation on the whole corpus of labeled gender and citizenship. Furthermore, as cross-validation is applied to prevent over-fitting, more than one ROC curve will be generated for each classification scenario. These will be combined using an algorithm called threshold averaging: A sample of the thresholds of all ROC curves is drawn. For each of these thresholds, one point from each ROC curve whose original threshold lied closest to the current threshold are taken and averaged to one ROC curve (Fawcett, 2006).

Overall, the two main classifiers we proposed, weighted soft-margin SVM classifier and supervised LDA, will be evaluated in the following. Furthermore, in order to gain better insights on the effect of topic models, the same SVM classifier will be applied to the output topics of unsupervised LDA. As performance indicators ROC (for binary classification), AUC, and also classification time duration will be considered to propose a model that can be used in future applications of user classification in Reddit.

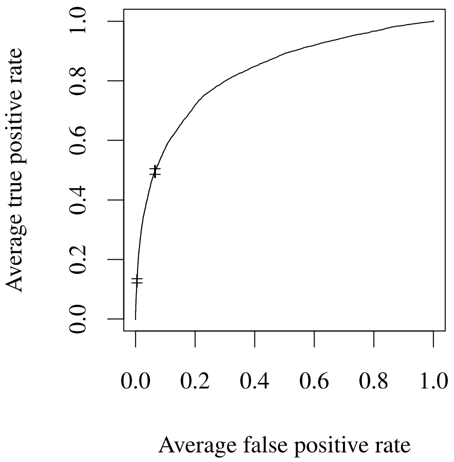

Figure 3. *ROC curve of SVM classification of gender without use of a kernel, based on Subreddits as features. The AUC of 82.9% is at a high level. The plot contains intervals on the graph to indicate the maximum deviation from the average across the folds.*

5.1. SVM Classifier

For the weighted soft-margin SVM classifier, one version with a radial base function (RBF) kernel and one without any kernel were chosen for comparison. The classification using the RBF kernel did not perform any better than a random classifier, as its AUC did not surpass the 50% threshold. This is consistent over the various versions of feature sets such as Reddits and Subreddits. Even the “simpler” binary classification of gender performed only at 50% AUC. Without the use of a kernel, classification of gender using Subreddits showed useful results with AUC of 82.9% (Figure 3).

5.2. SVM on LDA Topics

Although the SVM classifier already provided reliable results for the classification of gender, it might be possible to increase the performance even further by using other classifiers or different features. Instead of solely relying on the information provided by Subreddits, one can create topics with LDA and learn these with SVM classification. This approach is similar to the one Pennacchiotti and Popescu (2011a,b) applied for classification of political affiliation. Again, SVM classification using an RBF kernel did not perform better than random choice. Instead, the linear kernel provided better results.

As LDA creates an explicitly specified number of topics, we varied the number of topics to see how the AUC measure develops. For gender classification the performance depends heavily on the number of topics (Figure 4). With the use of topics it seemed to be even possible to discriminate citizenship inside the data. But these results were not promising as the highest average AUC (across the six folds) of only 53.8% was recorded at 350 topics. For gender, the highpoint was reached also at 350 topics but with a promising AUC of 87.3%. Therefore, an increase in performance was gained by moving from plain Subreddits to topics as features. Additionally the amount of time for training was reduced to a third over the method of SVM classification on Subreddits, making it more time-efficient.

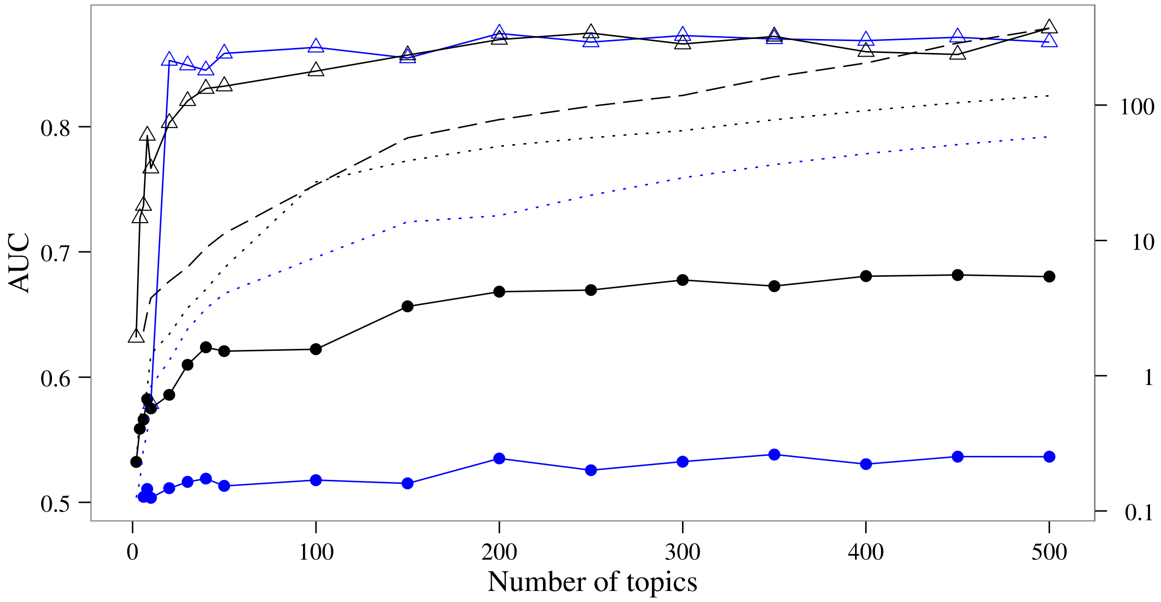

Figure 4. *AUC performance (continuous lines) for SVM-LDA (blue) and sLDA (black) for gender (Δ) and citizenship (•) with dependency on the number of topics. The dotted lines indicate the runtime in hours (with the y-axis on the right side) for gender and the dashed line for citizenship.*

5.3. sLDA

The performance of sLDA also relies on the number of topics. As *Figure* indicates, the larger the number of topics, the higher the AUC. Due to computational limitations, 500 was the maximum possible number of topics in this evaluation. Future work on this might be able to show whether the graph converges to a certain threshold (as Figure 4 suggests) or whether it reaches a peak and then declines from that point. Therefore, for classification of gender using Subreddits as features, the best performance was observed at 500 topics with an AUC of 87.9% which is also the best result of all previously presented classifiers. Similarly, classification of citizenship with Subreddits as features had its best AUC performance (68.2%) at 450 topics, indicating the best performance of any method to classify citizenship. Classification using Reddits as features did not perform better than a random classifier.

6. Discussion and Outlook

The evaluation showed a slight advantage for the algorithm of sLDA when classifying social gender, This advantage became very obvious when classifying citizenship (Figure 5). However, it should be noted that each of the proposed algorithms depends on several parameters that can influence the performance heavily as shown with the number of topics for sLDA. It may be that a better parameter configuration for the SVM classifier might result in a useful classification algorithm.

For comparison of the individual performances of the algorithms, Figure 4 also provides information on time duration. In this respect, the advantages of sLDA for gender classification become negligible. The LDA-SVM combination reaches a high level of AUC (85.3%) already at 20 topics while being 56% faster as its counterpart of sLDA at the same number of topics. But a comparable number of topics for sLDA with the same AUC is only achieved at 150 topics (85.7%) while taking 6,300% longer. For citizenship classification, the difference in AUC between LDA-SVM and sLDA are too big to consider any training and prediction times. Classification of citizenship using sLDA takes approximately three times longer than classification of gender using sLDA.

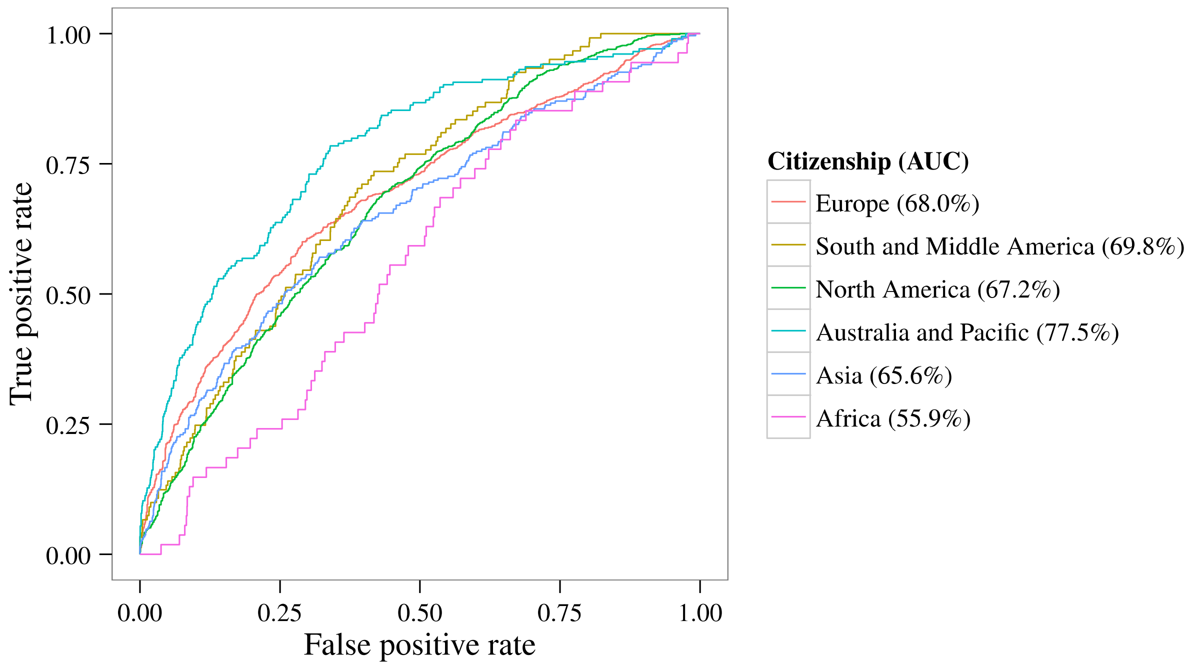

Figure 5. *ROC graphs for all classes of citizenship classified using sLDA with 500 topics on Subreddits. The ROC curves were acquired by plotting each class with the one-against-the-others method. This illustrates to what extent it is possible to identify continental citizenship of a user.*

There are some limitations of the Reddit corpus used in our study. Especially pre-tagging could involve some inconsistencies as firstly it only captures information on users that disclosed such information like being married to a "woman"/"man" or being "French" citizen. Secondly, this information – opposed to the approach of language-independent user classification of this article – only considers English comments. Creating a bigger amount of regular expressions in several languages could be one method that would decrease this inconsistency. Still the best approach might be to use manual tagging done by humans, although (Burger et al., 2011) showed that their classification algorithm could even classify with higher accuracy than humans.

Concerning limitations of classification models, the time-efficiency of sLDA is quite low compared to SVM classification or LDA combined with SVM classification. But for classifying citizenship it is advisable to use sLDA as the performance was significantly higher than with SVM or the LDA-SVM combination. One more time-efficient solution could be another Bayesian inference method called belief propagation which is supposed to be faster while also generating more accurate results than variational inference and also Gibbs sampling as claimed by (Zeng et al. 2013).

As support vector classification showed quite a high reliability for the prediction of gender, a combination of sLDA and SVM in one model could be a useful solution. It would – similar to sLDA – direct the topic generation and – similar to SVM – try to create a maximum margin between the classes. But this approach seems to be only helpful for classification of gender. For citizenship maximum-margin classification alone did not prove to be useful. One reason might be that it only “emulates” multi-class classification by using several binary classifications.

7. Conclusion

Our results show that it is possible to automatically classify gender and also citizenship of users on Reddit and therefore infringe their privacy, most probably without their awareness. This classification was performed based on users’ comments but without using full text analysis, thereby keeping the procedure language-independent at least to some extent. We also introduced a process of data extraction and pre-tagging to acquire a Reddit corpus which can be utilized for supervised learning. The presented methods conferred findings that approximately one third of the original extracted users could be annotated with the respective gender, and one fourth with citizenship divided into continental groups. Subsequently, SVM classification was applied. To improve classification performance, LDA was used as a generative model that attempts to extract intrinsic information from the user corpus by summarizing Subreddits or Reddits into topics. These topics could then be applied as features for SVM classification. Furthermore, sLDA was suggested as a version of LDA that directs the topic generation and also allows label prediction based on these topics.

Performance of these classifiers was evaluated using the previously acquired corpora of annotated gender and citizenship. As both corpora contained (heavily) skewed class distributions (e.g., less female than male users), the ROC together with the AUC have been chosen as measures for performance evaluation. Our experiments showed that a plain SVM without a kernel could classify gender with a high AUC performance. But for classification of citizenship, SVM did not prove to be an acceptable model. When SVM classification was based on topics that have been generated by LDA, the performance increased slightly, however, time-efficiency decreased drastically. Overall, sLDA performed best for gender and citizenship compared to the other models. Future work could include the investigation of further latent attributes and possibly also finer granularity of the citizenship attribute. But already our current results should increase user awareness of privacy risks when posting on apparently “safe” social media platforms such as Reddit.

References

Alexa Internet (2014). *reddit.com: Site Overview*. URL: http://www.alexa.com/siteinfo/reddit.com/ (visited 11/02/2014).

Baumann, A., Krasnova, H., Veltri, N. F., Ye, Y. (2015): "Men, Women, Microblogging: Where Do We Stand?" In: Thomas. O.; Teuteberg, F. (Hrsg.): Proceedings der 12. Internationalen Tagung Wirtschaftsinformatik (WI 2015), Osnabrück, 857–871

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

Burger, J. D. and J. C. Henderson (2006). "An Exploration of Observable Features Related to Blogger Age." In: *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pp. 15–20.

Fawcett, T. (2006). "An Introduction to ROC Analysis." *Pattern Recognition Letters* 27 (8), 861–874.

Ferguson, P., N. O'Hare, M. Davy, A. Bermingham, P. Sheridan, C. Gurrin, and A. F. Smeaton (2009). "Exploring the Use of Paragraph-Level Annotations for Sentiment Analysis of Financial Blogs." In: *Workshop on Opinion Mining and Sentiment Analysis (WOMAS 2009).*

Furnas, G. W., S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum (1988). "Information Retrieval Using a Singular Value Decomposition Model of Latent Semantic Structure." In: *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp. 465– 480.

Hand, D. J. and R. J. Till (2001). "A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems." *Machine Learning* 45 (2), 171–186.

Herring, S. (1996). "Two Variants of an Electronic Message Schema." *Pragmatics and Beyond New Series*, 81–108.

Herring, S. C. (2000). "Gender Differences in CMC: Findings and Implications." *Computer Professionals for Social Responsibility Journal* 18 (1).

Kaplan, A. M. and M. Haenlein (2010). "Users of the World, Unite! The challenges and opportunities of Social Media." *Business Horizons* 53 (1), 59–68.

Koh, M. (2013). *The 15 Best Discussions on Reddit Ever*. Thought Catalog.

Lakoff, R. T. (1972). *Language and Woman's Place*. Cambridge Univ Press.

Lotan, G., E. Graeff, M. Ananny, D. Gaffney, I. Pearce and D. Boyd (2011). "The Revolutions were Tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions." *International Journal of Communication* 5, 31.

MacKinnon, I. and R. H. Warren (2007). "Age and Geographic Inferences of the LiveJournal Social Network." In: *Statistical Network Analysis: Models, Issues, and New Directions*. Springer, pp. 176–178.

McAuliffe, J. D. and D. M. Blei (2007). "Supervised Topic Models." In: *Advances in Neural Information Processing Systems* 20, pp. 121–128.

Mukherjee, A. and B. Liu (2010). "Improving Gender Classification of Blog Authors." In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 207–217.

Netzer, O., R. Feldman, J. Goldenberg, and M. Fresko (2012). "Mine Your Own Business: Market-structure surveillance through text mining." *Marketing Science* 31 (3), 521–543.

Oelke, D., M. Hao, C. Rohrdantz, D. A. Keim, U. Dayal, L Haug, and H. Janetzko (2009). "Visual Opinion Analysis of Customer Feedback Data." In: *IEEE Symposium on Visual Analytics Science and Technology (VAST 2009)*. IEEE, pp. 187–194.

Osuna, E., R. Freund, and F. Girosi (1997). *Support Vector Machines: Training and Applications*. Technical Report. Massachusetts Institute of Technology.

Panyametheekul, S. and S. C. Herring (2003). "Gender and Turn Allocation in a Thai Chat Room." *Journal of Computer-Mediated Communication* 9 (1).

Pennacchiotti, M. and A.-M. Popescu (2011a). "A Machine Learning Approach to Twitter User Classification." *ICWSM* 11, 281–288.

Pennacchiotti, M. and A.-M. Popescu (2011b). "Democrats, Republicans and Starbucks Afficionados: User Classification in Twitter." In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 430–438.

Popescu, A.-M. and O. Etzioni (2007). "Extracting Product Features and Opinions from Reviews." In: *Natural Language Processing and Text Mining*. Springer, pp. 9–28.

Rao, D., D. Yarowsky, A. Shreevats, and M. Gupta (2010). "Classifying Latent User Attributes in Twitter." In: *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*. ACM, pp. 37–44.

Reddit Inc. (2011). *Who in the World is reddit? Results are in...* URL: http:// www.redditblog.com/2011/09/who-in-world-is-reddit-results-are-in.html/ (visited on 11/02/2014)

Reddit Inc. (2014). *Reddit API Documentation*. URL: http://www.reddit.com/dev/api (visited on 11/02/2014).

Redditor twentythree-nineteen (2013). *I downloaded 600,000 Reddit comments over a week*. URL: http://www.reddit.com/r/datasets/comments/1r76wp/i_downloaded_600000_reddit_comments_over_a_week/ (visited on 11/02/2014).

Schler, J., M. Koppel, S. Argamon, and J. W. Pennebaker (2006). "Effects of Age and Gender on Blogging." In: *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* 6, 199–205.

Tremayne, M. (2014). "Anatomy of Protest in the Digital Era: A network analysis of Twitter and Occupy Wall Street." *Social Movement Studies* 13 (1), 110–126.

Trudgill, P. (1972). "Sex, Covert Prestige and Linguistic Change in the Urban British English of Norwich." *Language in Society* 1 (02), 179–195.

Wang, Y.-C., M. Burke, and R. E. Kraut (2013). "Gender, Topic, and Audience Response: An analysis of user-generated content on Facebook." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, pp. 31–34.

Wikipedia (2014). *List of adjectival and demonymic forms for countries and nations*. URL: http://en.wikipedia.org/wiki/Adjectivals_and_demonyms_for_countries_and_nations/ (visited on 11/02/2014).

Yan, X. and L. Yan (2006). "Gender Classification of Weblog Authors." In: *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pp. 228–230.

Zeng, J., W. K. Cheung, and J. Liu (2013). "Learning Topic Models by Belief Propagation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (5), 1121–1134.

3.2 User Behavior and Mobile Devices

ARTICLE 17:

TO PHUB OR NOT TO PHUB: UNDERSTANDING OFF-TASK SMARTPHONE USAGE AND ITS CONSEQUENCES IN THE ACADEMIC ENVIRONMENT

Authors: Olga Abramova, Technische Universität Darmstadt, Darmstadt, Germany

Annika Baumann, Humboldt-Universität zu Berlin, Berlin, Germany

Hanna Krasnova, Universität Potsdam, Potsdam, Germany

Stefan Lessmann, Humboldt-Universität zu Berlin, Berlin, Germany

Proceedings: 25th European Conference on Information Systems (ECIS'17)

Publication year: 2017

Abstract

This study was inspired in part by calls for research to explore the ubiquitous phenomenon of phubbing in the academic environment. The goal of our study is to explore the phenomenon of phubbing and its consequences among students. Combining observations, questionnaires, quasi-experimental research design and focus groups interviews, our findings suggest that students phub a substantial amount of lecture time and often underestimate the effect this behavior has on their learning process. The quasi-experimental study shows that the number of times a student looks at a smartphone during the lecture is negatively related to the visual attention, while the total duration of smartphone use worsens the auditory attention. Follow-up analysis of the focus group interviews uncovers the causes of the phenomenon and possible preventive measures. The study thus contributes to a growing body of IS research on undesirable consequences of ICT use and provides implications for IS practitioners, simultaneously calling for a better solution of the problem commonly witnessed by the universities: the improvement of the educational process and student performance in the digital society.

Keywords: *Smartphones, Phubbing, ICT in Education, Multitasking*

1. Introduction

Increased availability of portable digital technologies made it a matter of course that information and communication technologies (ICTs) accompany our daily lives. Especially smartphones, with over 2 billion users worldwide, have become our everyday companion (Statista, 2016). Smartphones are used everywhere – at home, at work, at the playground, and even in the classroom when students are supposed to learn something new. In general, smartphones and other ICTs can be used to improve the education process, e.g. by providing better simulations and models (Condord Consortium, 2016), enabling learning (Coursera, 2016; Glovico, 2016) and facilitating better assessment (Kessler, 2010). In fact, lecturers experience the advantages of ICTs, reporting a positive impact on the educational process in 75% of the cases (Alex, 2007).

However, there is some evidence demonstrating that when it comes to learning, ICTs such as smartphones are a double-edged sword. If used inappropriately, devices in the classroom can cause distraction for learners (Fried, 2008; Jacobsen and Forste, 2011; Rosen et al., 2013; Gupta and Irwin, 2016) and their peers (Fried, 2008; Sana et al., 2013). Particularly smartphones, with 98% penetration rate among 18-24 aged people in developed countries (Nielsen, 2016), represent the major risk, since the combination of perceived ease of use, portability and a broad range of features and functionalities increase the chances that learners will engage in off-task behaviors (Wood et al., 2011).

Frequently referred to as "phubbing", ignoring the conversational partner in favour of one's own smartphone (Karadag et al., 2015; Chotpitayasunondh, 2016) has recently become a common behaviour among teenagers and adults, permeating child-parent communication (Radesky et al., 2014), work environment (Roberts, 2015) and romantic relationships (Coyne et al., 2011; McDaniel and Coyne, 2016; Roberts and David, 2016; Krasnova et al. 2016). In contrast to other settings, the educational environment often implies one-to-many communication, for instance in the form of front lecturing. This particularity of academic environments creates favourable ground for phubbing to be practiced. In fact, holding a lecture has become a real challenge for many professors who have to hold a lecture in front of learners, many of whom are glued to their glowing screens. Both academicians and teachers are puzzled by how to deal with the excessive smartphone use in the classroom: "*Even when I know I've created a well-structured and well-paced lesson plan, it seems as if no topic, debate, or activity will ever trump the allure of the phone*" (Barnwell, 2016). The most controversial is the fact that more than 80% of students (Berry, 2015) believe this to be an acceptable practice and perceive it as an established "new norm" (Chotpitayasunondh, 2016).

Against this background, the goal of our study is to explore the phenomenon of phubbing in the academic environment. In contrast to previous studies that often use survey data (e.g. Levine et al., 2007; Jacobsen and Forste, 2011; Rosen et al., 2013), we combine observations (Study 1), questionnaires (Study 1, 2), a quasi-experimental design (Study 2) and focus groups interviews (Study 3) to assess the prevalence of smartphone use during lectures, to investigate the patterns and motivations behind this behaviour and estimate the effect on educational outcomes. Moreover, comparing observed and self-reported data enables us to assess the magnitude of the estimation bias, when it comes to personal assessment of smartphone use.

The remainder of the paper is organized as follows. In the following section, we summarize related work and derive hypotheses that link personal study-unrelated smartphone use with the learning performance. In the next step, we present results of our qualitative study based on observations (Study 1), followed by the quasi-experiment (Study 2) and focus groups interviews (Study 3). Our results suggest that students spend substantial amount of time on their smartphones during the lecture. These findings justify further exploration of the effect of phubbing on visual and auditory attention during lectures. Analysis of the focus groups deepen our understanding of the causes of the phenomenon and allow us to derive possible preventive measures. Opportunities for future research and implications of our findings for IS research and practitioners are discussed in the concluding section.

2. Theoretical Background

Modern universities increasingly rely on ICTs to enable the construction of individual and collective knowledge (Holland and Judge, 2013). Since modern society is permanently online and permanently connected (POPC), the immediate and ubiquitous access to knowledge via the Internet has gotten so easy that our own knowledge (for example of some facts) plays a rather subordinate role (Vorderer, 2015). Following this new trend, the majority of universities provide students with permanent Internet access (Eduroam, 2016). While fostering learning, availability of free and unlimited Internet access also stimulates by-side smartphone activities during the class. We hypothesize that:

H1: Phubbing is a widespread phenomenon in the academic environment.

Several studies investigate the effect of the smartphone usage in the class on learning, linking the observed dynamics to the multitasking phenomenon (Table 1). In general, multitasking is defined as practicing more than one activity simultaneously (Pashler, 1994). In contrast to machines, humans are inclined to exhibit a "cognitive bottleneck" constraint in their decision-making (Welford, 1967), which shows up in slower performance of the secondary task (Levy and Paschler, 2001; McCann and Johnston, 1992; Pashler et al., 2008; Schumacher et al., 2001; Welford, 1952). Following this logic, smartphone use in the classroom for study-unrelated purposes is expected to negatively influence the academic success. According to research, short-term education outcomes are likely to suffer first (Table 1). For example, texting was found to have a detrimental effect on memorizing the lecture material (Ellis et al., 2010; Wood et al., 2011; Froese et al., 2012), although some studies have not confirmed this proposition (Rosen et al., 2011; Wood et al., 2011). A ringing phone during the class may affect not only the smartphone owner him/her-self but also fellow students, leading to lower scores on a comprehension test and missing corresponding information in the lecture notes (Shelton et al., 2009; End et al., 2009). Moreover, cell phone use has been shown to slow down the responses on the lexical decision task (Shelton et al., 2009).

Furthermore, Thornton et al. (2014) demonstrate that tasks with greater attentional and cognitive demands are extremely sensitive to any distractions, including the mere presence of the smartphone. Regarding the long-term academic performance (e.g., overall GPA), evidence on the influence of smartphone use remains mixed, as reflected in Table 1. Based on self-reported data, texting and engagement with Facebook when doing homework is negatively associated with college GPA, while for other activities, such as emailing, talking on the phone or using instant messaging no significant impact has been found (Junco and Cotton, 2012). Taken together, while research results remain mixed, there is growing evidence about the negative impact of smartphone use on the performance on tasks that require attention.

Learning theory (Dunn, 1983; Dunn, 1984; Reinert, 1976) suggests that there are three learning modalities: visual, auditory, and kinaesthetic/tactile abbreviated as VAK (Barbe et al., 1981). Fleming (1995) extended this model to VARK by adding the "reading/writing" construct. Multiple tests of the VAK/VARK model in past research suggest that the majority of students are multimodal (i.e. use several channels simultaneously) in their learning (Prithishkumar and Michael, 2014). In a traditional lecture setting, two forms are mainly prevalent: namely visual channel, achieved through lecture slides, and auditory channel, accomplished by the talk of the lecturer. We suggest that the use of smartphones during lectures affects students' attention through the aforementioned channels. In line with the past research, we approach phubbing via two dimensions:

1) quantitative (e.g. Rosen et al., 2011), defined as the number of times the smartphone is accessed; and 2) qualitative (e.g. Junco and Cotton, 2012), defined as the total duration of the phubbing session during the lecture.

We hypothesize that:

H2a. The number of phubbing sessions reduces visual attention.

H2b. The total duration of phubbing activities reduces visual attention.

H3a. The number of phubbing sessions reduces auditory attention.

H3b. The total duration of phubbing activities reduces auditory attention.

Study	**Device**	**Method**	**Measured SP activity**	**Performance-related variables (Relationship)**
Ellis et al. (2010)	SP	E	Texting	Lecture-based quiz score (-)
End et al. (2009)	SP	E	SP Rings	Comprehension test (-) Lecture notes (-)
Froese et al. (2012)	SP	E, S	Texting	Lecture-based quiz score (-)
Junco and Cotten (2012)	SP and other ICTs	S	FB use Texting Emailing Talking on SP Using IM	Overall college GPA (-) Overall college GPA (-) Overall college GPA (n.s.) Overall college GPA (n.s.) Overall college GPA (n.s.)
Rosen et al. (2011)	SP	E	Texting	Recall test (-/n.s.)
Shelton et al. (2009)	Phone	E	SP Rings	Quiz score (-) Response speed on lexical decision task (-)
Smith et al. (2011)	SP and other ICTs	E	SP conversation Texting	Memory Task (-) Memory Task (-)
Thornton et al. (2014)	SP	E	SP presence	Digit cancellation task (n.s.) Additive cancellation task (-)
Wood et al. (2011)	SP and other ICTs	E	Texting	Memory quiz (n.s.)

Table 1. Association between smartphone activities and learning performance: overview of selected studies. Note: SP-smartphone, E-experiment, S-survey, n.s. – not significant

3. Study 1: Understanding Real Behaviour and Self-Perceptions

In order to test our hypothesis H1, we conducted structured observations to assess the frequency of student phubbing activities during lectures in a purposive sample of bachelor students at one German university in summer term 2016. A lot of studies are conducted in either an experimental setting or use self-reports for data collection (Table 1). While these methods can be appropriate for several application areas, smartphone use may be different in artificial experimental setups as opposed to real environment. First, the habituation to the smartphone may be the reason of decreased control and poor recall. Second, classroom smartphone use may be perceived as socially undesirable (since it may signal disrespect to the lecturer), which may lead to underreporting. In this case, naturalistic observation which does "not interfere with the people or activities under observation" (Angrosino, 2005, p. 730), yields more reliable data. Observations are a standard method used across a variety of disciplines. This method is especially common in the context of smartphone use, since this activity is often conducted in public places and users often underestimate the time they engage in it. Indeed, a number of past studies use observation as a primary method of data collection to study smartphone use and addiction (e.g., Radesky et al. 2014; Thompson et al., 2013)

In the beginning of the observations, two observers took a seat in the middle of the lecture hall. Each of them selected three target seats while the lecture hall was still empty to be able to choose a student without selection bias; if the left-most seat stayed empty the person right from it was chosen. Observers monitored students seating in the range from row 7th to row 11th (median = 9th row). This was done to assure that we capture an "average student". The following parameters were recorded: gender, age, smartphone position in the beginning of the class, presence of other devices (e.g. notebook or tablet); start, end and type (e.g., browsing, texting) of each phubbing action as well as the reaction of neighbors.

At the end of the lecture, we asked the observed student to fill in a questionnaire in a paper form about his or her own estimated smartphone use and some demographic information, which allowed us to compare self-assessment with the observations' findings. The following questions were asked in a closed format: 1) For how long did you use your smartphone during this lecture? 2) For what purpose did you mainly use your smartphone during the lecture? 3) Could you follow the content of the lecture? 4) Did you get distracted by your smartphone? 5) If yes, how much? 6) Did your neighbor's behavior encourage you to use your smartphone? 7) Guess: How often did you use your smartphone during the lecture? 8) How strong was your interest in the topic of the lecture? 9) How did you find the lecture style of the professor? (to capture satisfaction with the style of lecture presentation), and 10) Why did you use your smartphone during the lecture?

3.1. Sample

We collected 60 observations (32 women vs. 28 men), which can be viewed as a rather balanced distribution considering the random choice of the target student. The average age in the sample is 20.5 years (min = 18 y.o, max = 27 y.o.). For the majority (more than 80%) it was the second semester at the university.

According to the Mann-Whitney U test, no significant differences were found between females and males in absolute phubbing time ($z=-0.326$, Prob $>|z|=0.744$) and relative phubbing time as a percentage of the lecture duration ($z=-0.652$, Prob $>|z|=0.514$). The subject of the lecture does not yield significant discrepancy in phubbing behavior based on Kruskal-Wallis Test with $\chi^2(2)$ =6.777, p=0.034 for absolute phubbing time and $\chi^2(2)$=5.311, p=0.07 for relative phubbing. Since the data significantly deviated from a normal distribution (Shapiro-Wilk test $p<0.05$ for both absolute and relative phubbing time), we used a non-parametric test.

Generally, the observations took place over the entire lecture duration. Therefore, the mean observation time accounted for 1 hour 22 minutes. Sometimes the observation had to be stopped earlier because of unexpected events: observed student has left or the lecture was finished earlier by the lecturer. 91.7% of the observed students had their smartphones already visible on the table from the very beginning and often started the class with their smartphones in their hand. For the majority (85%) the smartphone was the only device present on the table; three students had tablets and six students had laptops additionally on their table.

3.2. Activities: What Do Students Do on their Smartphones?

Our observations show that on average students practice phubbing activities about eight times during a lecture (mean=7.98; median=8). The least heavy users only accounted for two smartphone sessions, whereas the heaviest users made 21 queries into their smartphones. Since observers were sitting almost directly behind the target students, it was possible to note the specific uses of the smartphone. One single "phubbing session" often contained several actions, e.g. someone was browsing first, then got a message and continued to type a message. Table 2 shows the number and the share of students observed doing different activities on their smartphone during the lecture as well as the frequency and duration of phubbing actions.

The most interesting result shown here is that during lectures, texting and browsing are practiced by 91.7% and 90.0% of students respectively. A typical student from our sample devoted around 16 minutes of their smartphone time to messaging. Browsing or social network activities accounted for longer time periods and took around 20 minutes. Although the third favored action observed is looking at the screen in order to check the time or for updates (58.3% of observations), it takes only about 25 seconds on average. This can be explained by the rather small amount of time needed to complete these tasks. Focused reading was noticed among 38.3% of students with the average duration of about six minutes. Seven students (11.7%) used smartphones for playing games, spending around 4 minutes on entertainment. Activities such as photo shooting and reading were either related (e.g., photo of the professor's notes) or not related to the course (e.g., videotaping for snapchat). Taken together, phubbing activities not related to the learning process (i.e., texting, browsing, looking and playing) sum up to 40 minutes for an average student, thus occupying one-third of the lecture time.

Researchers also examined the surrounding of the observed students to see if any cascading behavior took place, i.e. students being triggered to use their smartphone by the smartphone use of other fellow students. In 23.3% of cases (14 observations) an observed person had no neighbors, whereas 22 students (36.7%) had peers sitting next to them. Of those, 30.0% of their fellow students (18 students) used their smartphone extensively, whereas 5.0% were not phubbing and for 1 observation it was not possible to get any results.

Use	Description	Frequency of action	Share of all actions (N=480)	Mean time in min
Looking	The student catches a quick glance at the screen for checking the time or if there is a new message without unlocking the phone.	79	16.5%	00:25
Texting	The student types something on the smartphone screen; usually a message at WhatsApp, Facebook or an e-mail.	234	48.8%	15:47
Browsing	The student swipes the finger from bottom to top of the smartphone screen to browse the internet; usually Facebook, Instagram, etc.	224	46.7%	20:10
Photo	The student takes a picture with the smartphone; either of the notes from the professor or of himself at Snapchat.	12	2.5%	00:54
Reading	The student scrolls down and carefully reads for example the news or study-related articles.	71	14.8%	05:47
Playing	The student taps on or swipes with his finger over the smartphone screen for playing a game.	22	4.6%	03:48
Calculator	The student uses the calculator application to solve an arithmetical problem.	7	1.5%	00:14
Other	Listening to the voicemail.	1	0.2%	00:03

Table 2. Ever-observed phubbing activities during the lecture. Note: mean time in minutes – average duration among all 60 observed students.

3.3. Questionnaire

After the observation, 56 of the monitored students filled out the questionnaire. The reason for the four missing responses are the cases when students left the class earlier or rejected the request.

22 respondents (39.3%) estimated the time phubbed during the lecture correctly (Figure 1), which we defined as being accurate to up to 5 minute difference. Surprisingly, two-third of them are "heavy phubbers" who spent more than half an hour with the device in total. This speaks for a conscious behavior, meaning that these students are in general aware how much they used their smartphone. While 21.4% of respondents were too self-critical and overestimated their phubbing behavior, other 41.1% of respondents definitely underestimated their smartphone use, among which 14.3% underestimated the time they used their smartphone for about 10-20 minutes. These differences in self-report vs. real behavior further support the importance of field data collection when it comes to capturing individual smartphone use, e.g. with the help of observations.

Responding to the question whether it was possible to follow the lecture (7-point Likert scale; 1=yes, 3=partly, 7=no), 10.7% agreed they could do so. 28.6% claimed that they were able to partly comprehend the material and 17.9% reported they could not follow the professor's presentation.

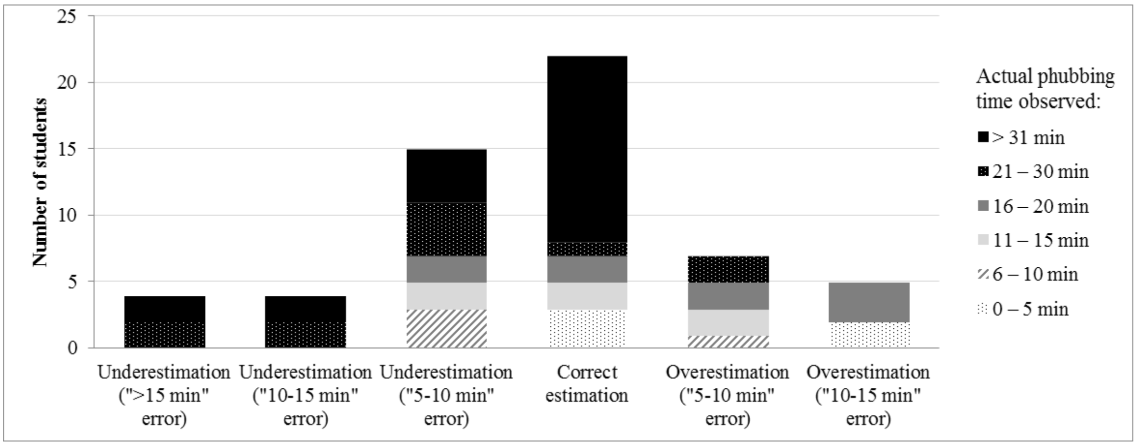

Figure 1. Students' estimation of the time phubbed vs. actual time phubbed.

The majority of respondents (55.4%) referred to the smartphone as a distraction during the lecture whereas 44.6% reported they were not disturbed. Those 31 students who felt distracted by their smartphone had to express to what extent they were distracted. Here, most students were only distracted a bit (around 50.0%) or barely (around 30.0%). However, the respondents did not shift the responsibility for their smartphone use to a neighbor: 50 of 56 respondents reported no influence on their smartphone behavior by the fellow students nearby.

For the next two questions, we controlled for the general attitude towards the subject and the satisfaction with the presentation style of the lecturer, which might have the potential to (at least) partly explain the phubbing behavior of respondents. Self-reported interest in the subject was low for the majority of respondents (60.8%), which can be partly attributed to the fact that mandatory courses were in the focus of our study. Presentation style of the lecturer was perceived as "rather good" or "good" in 37.5% of cases (see Table 3). To investigate whether the presentation style is related to the smartphone use we compared the average phubbing time for students who reported to be interested in the subject. We observe practically no difference in time spent on the smartphone regardless of the presentation style: both groups used their smartphone around 17 to 18 minutes. In case a student was not interested in the subject, we see a difference in the smartphone use: the average phubbing time was more than 25 minutes when the presentation was evaluated as good compared to 37 minutes when the presentation style was evaluated otherwise. As such, interest in the subject, hence own curiosity, seems to be a decisive factor.

Finally, we directly asked students about the reasons of their smartphone use during the lecture. The main reasons for phubbing according to the questionnaire are low satisfaction with the presentation style (60.7%), boredom (55.3%) and urgent message (51.8%). As already mentioned, there is a strong connection between the lecture style and boredom. The lower the satisfaction with the lecture style, the more boredom is reported, and the more easily respondents get distracted by their smartphone. These findings are in line with Lee et. al (2014) who state that smartphones are mainly used to get over boredom and so this is one of the main reasons why students engage in phubbing. All in all, the findings from Study 1 suggest that phubbing is common to the academic environment, thus confirming H1.

	High satisfaction with a presentation style		**Low satisfaction with a presentation style**	
	N (%)	**Average phubbing time**	**N (%)**	**Average phubbing time**
High interest in the subject	14 students (25.0%)	00:18:56	8 students (14.3%)	00:17:28
Low interest in the subject	7 students (12.5%)	00:25:12	27 students (48.2%)	00:37:01

Table 3. Average phubbing time and student assessment of the own interest in the course and the presentation style of the lecture.

4. Study 2: Phubbing and its Influence on Students' Performance

In Study 2 we empirically assessed whether the use of smartphones during lectures decreases the visual and auditory attention of students.

4.1. Quasi-experimental Design and Flow

For the quasi-experimental study (William et al., 2002), a 90-minutes lecture in Business Informatics at a large German university in the middle of the summer term 2016 has been chosen. The procedure included a two-part survey offered both in electronic and paper form. The first part of the survey was distributed at the beginning of the lecture with the notice that it was used to assess the quality of the lecture. It contained questions related to all former lectures regarding students' general satisfaction with the lecture ("How satisfied are with the lecture in general?"), the perceived usefulness of the lecture ("How useful do you find this lecture in general?"), the general learning growth ("How much do you usually learn in this lecture?"), the presentation style of the lecturer ("How do you find the presentation style of the lecturer?") and the general well-being and stress level of the student ("How do you feel?", adopted from Kross et al. (2013) and the motivation ("How motivated are you right now to study for this lecture?"). Questions were estimated on a scale ranging from zero to one hundred with latter being the best value. We used one-item scales for each question since keeping the questionnaire short was a priority considering the limited time frame of the lecture.

The second part of the survey took place at the end of the class and contained the same questions but related to the current lecture (e.g., "How much have you learned in today's lecture?"). We additionally asked questions with respect to smartphone use in terms of the general duration of smartphone activities ("How often have you used your smartphone during the lecture?") and frequency of smartphone sessions ("How many minutes you have used your smartphone during the lecture?") during the lecture. Furthermore, an open question was included where students had to state for what reason they used the smartphone ("Be honest: If you have used the smartphone during the lecture, why have you done this?"). Additionally, students had to state for what purpose ("How much of this time (in percentage, %) did you spend with one of the following applications? (Messaging, Social Networks, Non course-related use of Internet, Course-related use of the Internet, Games)" they used their smartphone. Finally, to check the relation between the surroundings and the person's intention to use a smartphone (Fried, 2008; Sana et al., 2013) we asked: "Have students in a direct proximity used the smartphone during the lecture?"

The educational outcomes – visual and auditory attention – were assessed by checking two pieces of information incorporated in the lecture and transmitted via only one channel. First, during the class a lecturer told a story about a Ph.D. student from Indonesia and further referred to the example 3-4 times repeating the country of origin. Auditory attention was measured by asking "Where does the former professor's Ph.D. student come from?" Second, on the slides which are usually designed in blue-white colours, a scheme in pink appeared to describe customer relationship management (CRM). This peculiarity was, however, intentionally not pointed out orally. We therefore asked later: "What colour did the CRM scheme have?" Both questions implied open answer and were then coded to binary variable (1- correct answer; 0 - false answer). At the end, we used student-selected unique identifiers to match both parts of the survey.

4.2. Sampling and Descriptive Statistics

A total of 77 respondents took part in our survey of the available 130 possible participants. 52% of the respondents in our sample are female. Almost all students (92.2%) reported that they used their smartphone during the lecture. Looking at the evaluation of student well-being and stress level at the beginning and at the end of the lecture we see only slight changes in well-being (the score of 69.8 in the beginning, and the score of 64.4 at the end) with a negative direction; whereas the stress level seems to be rather constant on average (the score of 57.1 at the beginning vs. the score of 56.6 at the end). Furthermore, in comparison with all former lectures, the present one was evaluated more positive in terms of its perceived usefulness (the score of 56.3 vs. 75.3), the satisfaction with the lecture (the score of 57.2 vs. 68.2), the presentation style of the lecturer (the score of 63.4 vs. 67.6) and the learning growth (the score of 49.8 vs. 56.8).

Regarding the smartphone use across gender during the lecture, we notice almost no difference in terms of frequency of smartphone use. However, when it comes to the duration of smartphone activities male students appear to spend more time with their smartphones compared to their female counterparts (see Figure 3, left). Asking for the purpose (why students used the smartphone), the survey responses are generally in line with the results of study 1. The reported purposes are messaging (42.3%), followed by non-course-related use of Internet (18.6%), course-related use of Internet (13.5%), social network use (12.4%), and games (2.0%) (see Figure 2, right).

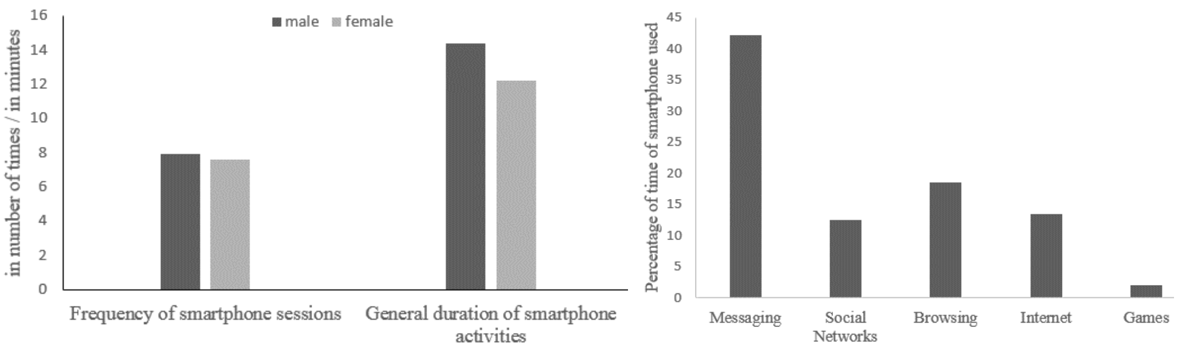

Figure 2. Frequency and duration of smartphone use per gender (left) and the purpose of smartphone use (right).

Additionally, almost all respondents reported that fellow students in their proximity used the smartphone during the lecture (72.4%), whereas only around 6.6% reported that they did not notice any phubbing next to them. However, 21% of respondents were not able to give an answer to this question.

Answering the question "If you have used the smartphone during the lecture, why have you done this?" respondents mainly reported texting as their main reason (43.5%), followed by boredom (18.8%) and concentration issues (14.9%). Some respondents also used the smartphone as a substitute for a watch (hence used it to check time) (6.9%), to read news (8.9%) and also for course-related activities (5%). Around 2% of the respondents also reported the use of their smartphone during the lecture as a result of it being a habit.

4.3. Results

To test the hypotheses proposed in Section 2, we did a logistic regression analysis for both visual and the auditory attention outcomes, since both variables were coded as binary (Table 4). Apart from that we also controlled for the well-being, fellow students' smartphone use, motivation, stress level and gender of the student, as well as the lecture evaluation variables (i.e., usefulness, presentation style, satisfaction, and learning growth).

Independent Variables	Visual Attention			Auditory Attention		
	Coefficient β	**Std. Error**	**Significance Level**	**Coefficient β**	**Std. Error**	**Significance Level**
Intercept	0.233	1.377	0.866	0.698	1.290	0.588
Frequency of Smartphone Use	**-0.186***	0.089	0.036	0.087	0.060	0.149
Duration of Smartphone Use	0.032	0.035	0.365	**-0.080***	0.037	0.031
Stress	-0.001	0.012	0.903	-0.011	0.012	0.346
Motivation	-0.022	0.018	0.223	0.024	0.017	0.158
Usefulness	-0.012	0.019	0.541	0.003	0.017	0.858
Presentation Style	0.007	0.020	0.706	-0.004	0.017	0.829
Satisfaction	0.014	0.026	0.574	-0.023	0.023	0.331
Learning Growth	0.001	0.014	0.939	0.011	0.014	0.442
Gender	-0.575	0.493	0.243	0.595	0.485	0.220
Fellow Student Use of Smartphone	0.051	0.395	0.897	-0.583	0.429	0.174
Nagelkerke Pseudo R-squared	0.187			0.240		

Table 4. Results for regression coefficients, standard error and significance of the logistic regression ($p < 0.05$).*

We observe that the frequency of smartphone use significantly reduces the visual attention. This indicates that the smartphone interactions that take place during the lecture – even if they are only brief –do have a negative influence on how well a student can follow the slides presented during the lecture (H2a confirmed). The coefficient for the total duration of the smartphone use was statistically insignificant (H2b rejected).

In contrast, auditory attention is negatively influenced by longer smartphone sessions. In other words, the more time respondents spend with the smartphone the less they are able to correctly memorize the content presented orally (H3b confirmed). No significant impact of the frequency of smartphone use on the auditory channel has been found (H3a rejected).

In summary, results of the logistic regressions show that the number of times a student looks at a smartphone during a lecture is negatively related to his or her visual attention. It is reasonable because the number of times one is distracted from the lecture slides results in one missing some visual information. Second, the total amount of time a student devotes to a smartphone is negatively related to auditory attention. As such, the longer a person uses the smartphone, i.e. the deeper the involvement with the smartphone is, the less attentively one is able to listen to the lecturer.

5. Study 3: Using Focus Groups to Explore Reasons of Phubbing Among Students and Opportunities to Reduce It

In order to gain better understanding into students' phubbing behavior, its antecedents and reactions, two focus group interviews were conducted at one German university in November 2016. This method allows researchers to "tease out the strength of participant's beliefs and subtleties about the topic that may be missed in individual interviews" (Campbell, 1988). Based on the literature overview and discussion among the authors, the following three items targeting phubbing in the academic context were generated and included in the protocol:

1) Do you check the smartphone or entertain yourself with the smartphone during lectures? What could be the reasons for this behavior?
2) In your opinion, how do smartphone activities during a lecture influence the performance? Does checking the smartphone help you to relax quickly? Or do you feel negative consequences of distraction, e.g. it is difficult to follow the lecture?
3) Do you think it is possible to reduce phubbing during lectures? Why? If so, how is it possible?

Two focus group interviews were organized, with 8 students (2 males and 6 females) in the first group and 6 female students in the second one. For analytical purposes, both focus group results were combined into one dataset. According to the short questionnaire completed in the beginning of the discussion, the majority (78.6%) of respondents study Business Informatics and are 26 to 30 years old; all others (21.6%) study Business and are 21 to 25 years old. All respondents have a smartphone; however, half of the sample got it after their 20th birthday. Five respondents (35.7%) have owned the device since they are 16 to 20 years old, and two respondents got used to smartphones as teenagers as they were 11 to 15 years old. Most frequently smartphones are used for emails and social media (64.3%) and most of the respondents (57.1%) check it several times an hour.

Our first research question aimed to elicit the prevalence of phubbing during lectures. Our data suggests that it is common that students use their smartphone during the lecture (P2.6: "*Of course I do it, I mean, sometimes it's more, sometimes it's less*"), with two exceptions (P1.5. and P1.3) where a radical way to preclude this was chosen: P1.3 "*…I live at the campus … so I just left the phone at home for two hours so that I don't get in the situation I want to take it out*". When specifying the reasons, it is possible to differentiate between the kickoff and protracted absorption triggers. Initial unlocking of the smartphone is usually rooted in concentration problems (P2.6 "*very often I'm off…I'm just not concentrating anymore but I'm really trying not to do it",* P2.4 *"it is just about the self-control which is not that present sometimes*") or the sense of boredom during a lecture (P1.4 "*if the lecture is not so interesting…*" (P1.8, P2.3- the same). Apart from content, the presentation style matters as pointed out P1.7, "*there is an interactive kind of lecture that doesn't really give you the chance to look at the smartphone that often and there is this … ehm … frontal version of lecture where you .. like disconnected from the teacher*", which is in line with our findings from Study 1. In contrast, lasting phubbing may be arranged in advance illustrated by P2.6 : "*it has to do with private things I' organizing like…ahm…meeting friends or checking what I have to…to buy in the evening (laughing)….or ahm….like…what other things have to be …it's not really entertainment…*" Similarly, P2.1 said "*it's more like what I have planned... If I have thing very urgent … or something I have been thinking over a whole day: I need to write that person, I need to write this, I need to write that. …it's just because I have things that I need to do on my phone, then…it doesn't matter if it [lecture] was interesting or boring*", disputing the importance of style and content of the lecture. Even if enduring phubbing was not intended, after a quick check, students are swamped by the multifunctionality of the device and permanent updates leading to absorption with the smartphone, summarized by P2.4: *"…you switch on your phone and then…oh… I have a message and* *then I'm tagged somewhere on a new picture or let's take a look who is this so (laughing)…so yeah…it really can be such a sequence of unwanted actions actually…*"

Referring to the second research question about the influence of phubbing during lectures on performance, students admit decreasing attention and debunk the myth about multitasking. For example, P1.3 reported: "*I think I pay less attention to the lecture…I cannot listen if I am writing a text message, you think you can but actually you can't*". Similar ideas are expressed by P1.2. ("*you lose information*"), P2.2. ("*cannot keep up with the lecture anymore*"), P1.7 (…"*can't focus on the contents that are presented, in the moment you are distracted…*"), P2.5. ("*the performance goes down.. like.. definitely goes down*"). However, some respondents claim that phubbing won't influence the final grade for the class because they will catch up later. For example, P1.3 suggests "*…if I don't pay full attention in the lecture I know I have to go through the information again when I learn for the exam*" or P2.6. "*…and you have to do more at home (laughing)*". In general, as P1.7. mentioned: "*a negative effect in inefficiency! …. that leads to the consequence that you have to focus on the content another time*".

Finally, we asked participants to reflect on possible ways to reduce phubbing during lectures. "*If smartphone is on the table already (smiling), it's very easy to have a quick look in your messages, and so on*" responded P2.6 and therefore it was proposed to leave smartphones in the bag (P2.6, P1.4, P1.7) or to switch on the flight mode because it is "*a good solution to not receive anything...not to be distracted by push messages*", as noticed P1.7. The majority agrees that "*restriction won't work well*" as P2.1 said. At the same time, P2.4 explains that even in the absence of the signal a student "*finds something [P2.6 is nodding her head] ...he can draw [laughing]… just use old-school methods to entertain yourself...there are plenty of [laugh]*". P2.4. experienced that students "*just substituted it [smartphone] with their laptops… they just did the same thing with Candy Crush and whatever stuff on the laptops*". Instead, P2.1 and P1.4 encourage increasing awareness and "*tell them what effect it would have*" (P2.1). However, students find the best way to reduce phubbing is to "*fight fire with fire*", namely, to develop a smartphone application and thus "*integrate functions of the smartphone into the whole lecture, for example surveys*" (P1.7). Similarly, P1.2 proposed: "*I thought about using questionnaires…so that everyone in the lecture has to seek answers a,b,c,d like in the "Who will be a millionaire?*". This will give "*an instant feedback on the topic of the lecture...how many [students] understood…*" Thus participants perceived a need for more interaction between a lecturer and students during the class which, accounting for the ubiquitous addiction to devices, could be established through the smartphone.

6. Discussion, Implications and Concluding Remarks

This work demonstrates that phubbing is common in academic settings. Three studies showed that students use their smartphone a substantial amount of lecture time and may underestimate the effect this behavior has on the learning process. The results of study 1 show that study-unrelated activities like texting, browsing, looking at the screen and playing take about one-third of the lecture time. Regarding study-related activities on the smartphone, e.g. looking up an unknown definition or using calculator, students allocate 1% of time. However, the majority of respondents are aware of the time lost, although some "heavy users" strongly underestimate time spent with the smartphone with a more than 10-15 minute error. Almost one third of the observed students claimed they were able to follow the presented material only partly, thus admitting the diminishing concentration, while more than 50% answered that they able to (partly) follow the lecture.

The results of quasi-experimental study 2 suggest significant adverse effects of phubbing during lecture on attention and learning. As such, the number of times a person looks at the smartphone screen is negatively related to visual attention. This effect seems to take place because frequent distraction from the lecture slides naturally leads to the loss of the visual information. The amount of time a student devotes to the device is also negatively related to his or her auditory attention. Our argument is that long smartphone sessions usually imply deeper involvement with the activity which means students listen to the lecture less carefully.

The results of study 3, designed as focus groups interviews, in combination with surveys embedded in study 1 and study 2 offer insights into why students practice phubbing, how they perceive the effects of phubbing, and whether it is possible to prevent it. As such, low interest in the lecture, low satisfaction with the presentation style of the lecturer as well as self-control issues are the main reasons for off-task smartphone activities. Although negative effects on instant educational outcomes were admitted, the majority of respondents believe phubbing at the lecture does not influence the long-term outcomes, namely the exam grade, since they plan to go through the material once again. To prevent the excessive smartphone engagement, it is recommended not to put the device on the table leaving it in the bag or switching on the flight modus in order not to be distracted by constantly incoming messages and newsfeed updates.

Our findings have implications for IS practitioners mainly targeting mobile app providers and smartphone producers. To the best of our knowledge, there exist only few applications addressing the phubbing issue at school, at work or at home (Flipdapp.co, 2017; Xerofone.com, 2017). Narrowing the perspective to the learning environment, students (study 3) believe the best way to solve the problem is to create a smartphone application that allows to give an immediate feedback to the lecturer on the material understood and thus helps to keep attention (Dyer, 2016). Another opportunity is an application that monitors phubbing activities and makes students aware of the total amount of time spent inefficiently during learning (Goldman, 2015). Moreover, raising awareness about the scale of phubbing in the educational context may be desirable.

This study was inspired in part by calls for research to explore the ubiquitous phenomenon of phubbing in the academic environment, previously studied in the romantic (e.g., McDaniel and Coyne, 2016; Roberts and David, 2016; Krasnova et. al., 2016) and family context (Radesky et al., 2014). Our aim was to understand the phubbing behavior of learners in the academic context, as well as to gain a better understanding of its antecedents and consequences. However, the current study comes with limitations that open exciting venues for future research. First, our investigation can be extended to a broader range of subjects and type of classes to include seminars and tutorials, thus increasing the reliability of the results. Moreover, our findings are especially valid for academic institutions that have large classes and a high level of smartphone adoption among students. Finally, to extend our results, a more comprehensive model describing phubbing influence on learning can be tested in future studies.

Acknowledgements

We would like to thank Sarah Mittelstaedt and Maria Herrmann, Master students at the University of Potsdam, for their support with this research.

References

Alex, J. (2007). *Does Using Technology Help Students Retain Information They Learn?* URL: http://ed.psu.edu/pds/teacher-inquiry/2007/alexjinquiry0607.pdf (visited on 12/03/2016).

Angrosino, M. V. (2005). *Recontextualizing observation: Ethnography, pedagogy, and the prospects for a progressive political agenda*. In: Denzin, N. K., and Y. S. Lincoln, *The SAGE Handbook of Qualitative Research*. Thousand Oaks: SAGE Publications.

Barbe, W. B. and Milone Jr, M. N. (1981). What We Know About Modality Strengths. *Educational Leadership*, 38(5), 378-80.

Barnwell P. (2016). *Do Smartphones Have a Place in the Classroom?* URL: http://www.theatlantic.com/education/archive/2016/04/do-smartphones-have-a-place-in-the-classroom/480231/ (visited on 12/11/2016).

Berry, M. J. and A. Westfall. (2015). Dial D for distraction: The making and breaking of cell phone policies in the college classroom. *College Teaching*, *63*(2), 62-71.

Campbell, D., 1988. *Methodology and Epistemology for Social Science*. Chicago, University of Chicago Press.

Chotpitayasunondh, V. and K. M. Douglas (2016). How "phubbing" becomes the norm: The antecedents and consequences of snubbing via smartphone. Computers in Human Behavior, 63, 9-18.

Condord Consortium (2016). *The Concord Consortium – Revolutionary digital learning for science, math and engineering*. URL: https://concord.org/ (visited on 12/03/2016).

Coursera (2016). *Online Courses From Top Universities*. URL: https://www.coursera.org/ (visited on 12/11/2016).

Coyne, S., Stockdale, L., Busby, D., Iverson, B. and D. Grant (2011). " 'I luv u :)!': A descriptive study of the media use of individuals in romantic relationships." *Family Relations* 60 (2), 150-162.

Dunn, R. (1983). "Learning style and its relation to exceptionality at both ends of the spectrum." *Exceptional Children* 49, 496-506.

Dunn, R. (1984). "Learning style: State of the scene." *Theory Into Practice* 23, 10-19.

Dyer K. (2016). *Take Three! 55 Digital Tools and Apps for Formative Assessment Success* URL: https://www.nwea.org/blog/2016/take-three-55-digital-tools-and-apps-for-formative-assessment-success/ (visited on 12/11/2016).

Eduroam (2016). *Eduroam*. URL: https://www.eduroam.de/eduroammap/ (visited on 12/03/2016).

Ellis, Y., Daniels, B. and A. Jauregui (2010). The effect of multitasking on the grade performance of business students. *Research in Higher Education Journal 8*, 1.

End, C. M., Worthman, S., Mathews, M. B. and K. Wetterau (2009). "Costly cell phones: The impact of cell phone rings on academic performance." *Teaching of Psychology* 37 (1), 55-57.

Fleming, N. D. (1995). "I'm different; not dumb. Modes of presentation (VARK) in the tertiary classroom. In Research and Development in Higher Education." In: *Proceedings of the 1995 Annual Conference of the Higher Education and Research Development Society of Australasia (HERDSA)*. HERDSA, Vol. 18, pp. 308-313.

Flipd (2017). Flipd – Remove Distractions. URL: http://flipdapp.co/ (visited on 04/21/2017).

Fried, C. B. (2008). "In-class laptop use and its effects on student learning." *Computers & Education* 50 (3), 906-914.

Froese, A. D., Carpenter, C. N., Inman, D. A., Schooley, J. R., Barnes, R. B., Brecht, P. W. and J. D. Chacon (2012). "Effects of classroom cell phone use on expected and actual learning." *College Student Journal* 46 (2), 323-332.

Glovico (2016) *Fairtrade Online Sprachschule*. URL: http://www.glovico.org (visited on 12/03/2016).

Goldman J. (2015). *6 Apps to Stop Your Smartphone Addiction.* URL: http://www.inc.com/jeremy-goldman/6-apps-to-stop-your-smartphone-addiction.html. (visited on 12/11/2016).

Gupta, N. and J. D. Irwin (2016). "In-class distractions: The role of Facebook and the primary learning task." *Computers in Human Behavior* 55, 1165-1178.

Holland, C. and M. Judge (2013). *Future Learning Spaces: The Potential and Practice of Learning 2.0 in Higher Education*.

Jacobsen, W. C. and R. Forste (2011). "The wired generation: Academic and social outcomes of electronic media use among university students." *Cyberpsychology, Behavior, and Social Networking* 14 (5), 275-280.

Junco, R. and S. R. Cotton (2012). "No A 4 U: The relationship between multitasking and academic performance." *Computers & Education* 59 (2), 505-514.

Karadag, E., Tosuntas, S. B., Erzen, E., Duru, P., Bostan, N., Mizrak, S., Culha, I. and B. Babadag (2015). „Determinants of phubbing, which is the sum of many virtual addictions: A structural equa-tion model." *Journal of Behavioral Addictions*, 4 (2), 60-74.

Kessler, S. (2010). *8 Ways Technology Is Improving Education*. URL: http://mashable.com/2010/11/22/technology-in-education/#NNd3ox5GqkqC (visited on 12/02/2016).

Krasnova, H., Abramova, O., Notter, I. and A. Baumann (2016). “Why Phubbing is Toxic for your Relationship: Understanding the Role of Smartphone Jealousy Among “Generation Y” Users.” In: *European Conference on Information Systems.*

Kross, E., Verduyn, P., Demiralp, E., Park, J., Lee, D. S., Lin, N., Shablack, H., Jonides, J. and O. Ybarra (2013). “Facebook use predicts declines in subjective well-being in young adults.” *PloS one* 8 (8).

Lee, U., Lee, J., Ko, M., Lee, C., Kim, Y., Yang, S., Yatani, K., Gweon, G., Chung, K.-M. and J. Song (2014, April). “Hooked on smartphones: an exploratory study on smartphone overuse among college students.” In: *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, pp. 2327-2336.

Levine, L. E., Waite, B. M. and L. L. Bowman (2007). “Electronic media use, reading, and academic distractibility in college youth.” *CyberPsychology & Behavior* 10 (4), 560-566.

Levy, H. and H. Paschler (2001). “Is dual-task slowing instruction dependent?” *Journal of Experimental Psychology: Human Perception and Performance* 27 (4), 862–869.

Li, Y., Medwell, J., Wray, D., Wang, L., & Xiaojing, L. (2016). Learning Styles: A Review of Validity and Usefulness. *Journal of Education and Training Studies*, 4(10), 90-94.

McCann, R. and J. C. Johnston (1992). “Locus of the single-channel bottleneck in dual-task interference.” *Journal of Experimental Psychology: Human Perception and Performance* 18, 471–485.

McDaniel, B. T. and S. M. Coyne (2016). “’Technoference’: The interference of technology in couple relationships and implications for women’s personal and relational well-being.” *Psychology of Popular Media Culture* 5 (1), 85.

Nielsen (2016). Millennials are Top Smartphone Users. URL: http://www.nielsen.com/us/en/insights/news/2016/millennials-are-top-smartphone-users.html (visited on 12/02/2016).

Pashler, H. (1994). "Dual-task interference in simple tasks: data and theory.” *Psychological Bulletin* 16, 220–244.

Pashler, H., Harris, C. R. and K. H. Nuechterlein (2008). "Does the central bottleneck encompass voluntary selection of hedonically based choices?” *Experimental Psychology* 55 (5), 313–321.

Prithishkumar, I. J., & Michael, S. A. (2014). Understanding your student: Using the VARK model. Journal of postgraduate medicine, 60(2), 183.

Radesky, J. S., Kistin, C. J., Zuckerman, B., Nitzberg, K., Gross, J., Kaplan-Sanoff, M., Augustyn, M. and M. Silverstein (2014). “Patterns of mobile device use by caregivers and children during meals in fast food restaurants.” *Pediatrics* 133 (4), e843-e849.

Reinert, H. (1976). “One picture is worth a thousand words? Not necessarily.” *Modern Language Journal* 60, 160-168.

Roberts, J. A. (2015). *Did you get phubbed by your boss?*. URL: http://edition.cnn.com/2015/10/06/opinions/roberts-boss-phubbing/ (visited on 12/03/2015).

Roberts, J. A. and M. E. David (2016). “My life has become a major distraction from my cell phone: Partner phubbing and relationship satisfaction among romantic partners.” *Computers in Human Behavior* 54, 134-141.

Rosen, L. D., Carrier, L. M. and N. A. Cheever (2013). “Facebook and texting made me do it: Media-induced task-switching while studying.” *Computers in Human Behavior* 29 (3), 948-958.

Rosen, L. D., Lim, A. F., Carrier, L. M. and N. A. Cheever (2011). “An empirical examination of the educational impact of text message-induced task switching in the classroom: Educational implications and strategies to enhance learning.” *Psicología educativa* 17 (2), 163-177.

Russell, M., Price, R., Signal, L., Stanley, J., Gerring, Z., & Cumming, J. (2011). What do passengers do during travel time? Structured observations on buses and trains. Journal of Public Transportation, 14(3), 7.

Sana, F., Weston, T., & Cepeda, N. J. (2013). “Laptop multitasking hinders classroom learning for both users and nearby peers.” *Computers & Education* 62, 24-31.

Schumacher, E. H., Seymour, T. L., Glass, J. M., Fencsik, D. E., Lauber, E. J., Kieras, D. E. and D. Meyer (2001). “Virtually perfect time sharing in dual-task performance: uncorking the central cognitive bottleneck.” *Psychological Science* 12 (2), 101–108.

Shelton, J. T., Elliott, E. M., Eaves, S. D. and A. L. Exner (2009). “The distracting effects of a ringing cell phone: An investigation of the laboratory and the classroom setting.” *Journal of Environmental Psychology* 29 (4), 513-521.

Smith, T. S., Isaak, M. I., Senette, C. G. and B. G. Abadie (2011). “Effects of cell-phone and text-message distractions on true and false recognition.” *Cyberpsychology, Behavior, and Social Networking* 14 (6), 351-358.

Statista (2016). Number of smartphone users worldwide from 2014 to 2020 (in billions). URL: https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/ (visited on 12/03/2016).

Thompson, L. L., Rivara, F. P., Ayyagari, R. C. and Ebel, B. E. (2013). "Impact of social and technological distraction on pedestrian crossing behaviour: an observational study." *Injury Prevention* 19(4), 232-237.

Thornton, B., Faires, A., Robbins, M. and E. Rollins (2014). “The Mere Presence of a Cell Phone May be Distracting.” *Social Psychology* 45 (6), 479-488.

Vorderer, P. (2015). „Der mediatisierte Lebenswandel, Permanently online, permanently connected,“ *Publizistik, Springer Fachmedien Wiesbaden*, pp. 259-276.

Welford, A. T. (1952). “The “psychological refractory period” and the review of high speed performance: a review and theory.” *British Journal of Psychology* 43, 2–19.

Welford, A. T. (1967). “Single-channel operation in the brain.” *Acta Psychologica* 27, 5–22.

William R., Shadish, Cook, T. D. and Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Wadsworth Cengage learning.

Wood, E., Zivcakova, L., Gentile, P., Archer, K., De Pasquale, D. and A. Nosko (2012). “Examining the impact of off-task multi-tasking with technology on real-time classroom learning.” *Computers & Education* 58 (1), 365-374.

Xerofone.com (2017). This is Xerofone – We can manage smartphones and limit their us. URL: http://xerofone.com/ (visited on 04/21/2017).

ARTICLE 18:

WHY PHUBBING IS TOXIC FOR YOUR RELATIONSHIP: UNDERSTANDING THE ROLE OF SMARTPHONE JEALOUSY AMONG "GENERATION Y" USERS

Authors: Hanna Krasnova, Universität Potsdam, Potsdam, Germany

Olga Abramova, Technische Universität Darmstadt, Darmstadt, Germany

Isabelle Notter, Universität Bern, Bern, Switzerland

Annika Baumann, Humboldt-Universität zu Berlin, Berlin, Germany

Proceedings: 24th European Conference on Information Systems (ECIS'16)

Publication year: 2016

Abstract

Coined as "phubbing", excessive use of smartphones in the romantic context has been shown to represent a barrier to meaningful communication, causing conflict, lowering relationship satisfaction, and undermining individual well-being. While these findings project a dire picture of the future of romance, the mechanisms behind the detrimental influence of partner phubbing on relationship-relevant markers are still little understood. Considering prior evidence that partner phubbing leads to the loss of exclusive attention towards the other party, we argue that these are rather the feelings of jealousy partner phubbing is triggering that are responsible for the negative relational outcomes. Based on the analysis of qualitative and quantitative responses from "generation Y" users, we find that partner phubbing is associated with heightened feelings of jealousy, which is inversely related to couple's relational cohesion. Moreover, jealousy plays a mediating role in the relationship between partner's smartphone use and relational cohesion, acting as a mechanism behind this undesirable link. Challenging the frequently promoted euphoria with regard to permanent "connectedness", our study contributes to a growing body of IS research that addresses dark sides of information technology use and provides corresponding implications for IS practitioners.

Keywords: *Smartphones, Social Media, Phubbing, Relational Cohesion, Jealousy*

"The first couple of weeks he was on his phone 24/7. I assumed it was just the novelty of having a smartphone for the first time and I didn't think anything of it. But it never stopped. All of "our" time just twisted into him being on his phone. I was practically begging for his attention. I'd try to have deep conversations; he'd be on Reddit. I'd try snuggling and being cute; he'd be playing Heartstone. [...]. We can't have a quiet evening together [...] without his phone competing for his attention. I'm lonely and depressed." (MissHurt, 2015)[21]

1. Introduction

We are in a coffee shop and we observe: *A couple walks in. She already has her smartphone in the hand. They sit down on opposite sides of the table. While he grabs some food for both of them, she starts to immediately focus on her smartphone, constantly scrolling and swiping. When he returns she stops using it for just a minute. Once they start drinking their coffee, she keeps on interacting with her mobile device. He gets visibly bored and also takes out his smartphone, possibly to just have something to do. She notices and passes him her smartphone to show him something. When he returns her smartphone, she continues using it for almost 30 minutes straight. Meanwhile he goes through a routine to pick up his smartphone for a few minutes only to put it away for a short time and to grab it again, seemingly bored. They rarely talk to each other while looking at their smartphones. After about an hour they leave together. When he puts on his jacket, she still keeps looking at her smartphone.*

With around 3.4 billion users worldwide (Ericsson Mobility Report, 2015), it is not surprising that smartphones are increasingly permeating our daily routines: We use them on the railway station waiting for the train, in the bus that brings us home. We use them when we meet friends, when driving cars (Smith, 2015), or crossing a busy road on a pedestrian walkway (Hatfield and Murphy, 2007). For many, smartphones are the first thing they touch when waking up, and the last one they look at before going to sleep (Cisco, 2014). Fueled by the widespread interest in Social Media apps (Salehan and Negahban, 2013), using smartphones is fun, useful, informative, and highly addictive (e.g. Jung, 2013). In fact, studies show that 81 percent of users keep their smartphones nearby for the entire day and check it 110 times per day on average (Woollaston, 2013).

Indisputably, the widespread adoption and usage of smartphones has changed our lives. However, the nature of these transformations is still ambiguous. Some studies report the positive influence of smartphones in professional environments such as healthcare coordination (Wu et al., 2011, Whitlow et al., 2014; Wickersham et al., 2015), infrastructure monitoring (Mohan et al., 2008, White et al., 2011), or simply emphasize their value in promoting socialization with geographically distant individuals (Smith, 2015; Amplitude Research, 2013). At the same time, another stream of research stresses the harmful consequences of smartphone interference across a variety of communication contexts, including face-to-face conversations (McDaniel and Coyne, 2016), parents-child interaction (Devitt and Roker, 2009), work-related management activities (Roberts, 2015) and educational processes (Ling, 2000; Campbell, 2005). Among these findings, the insights into the damaging role of smartphones in the romantic context are particularly alarming.

Indeed, coined as "phubbing", snubbing the romantic partner when using the smartphone in his or her company has been shown to cause conflict, lower relationship satisfaction, and individual well-being (McDaniel and Coyne, 2016; Roberts and David, 2016). While these findings project a dire picture of the future of romance and family structures, the mechanisms behind the detrimental influence of partner phubbing on relationship-relevant markers is still little understood. As of now, existing research suggests that smartphones may represent a barrier to meaningful communication, provoking feelings of constant

[21] This quote has been edited for style to improve readability. Original can be found at: https://www.reddit.com/r/TwoXChromosomes/comments/3lmz1h/i_know_a_lot_of_things_can_create_problems_in/

interruption, disrespect (Duran et al., 2011, Tertadian, 2012) and irritation (Theiss and Solomon, 2006; Roberts and David, 2016). However, the mechanism behind these negative resentful reactions remains uncovered. To fill this gap and considering that partner phubbing inevitably leads to the loss of exclusive attention towards the other party, we argue that these are rather the feelings of *jealousy* partner phubbing is triggering that are responsible for the negative relational dynamics reported in past research. Indeed, defined as "*a protective reaction to a perceived threat to a valued relationship, arising from a situation in which the partner's involvement with an activity and/or another person is contrary to the jealous person's definition of their relationship*" (Bevan and Samter, 2004, p. 15), jealousy incorporates loss of exclusive attention as one of its major premises (Bauminger, 2010; Tov-Ruach, 1980). Negative in its essence, jealousy has commonly been associated with such undesirable relational outcomes as expressions of aggression and conflict (Guerrero et al., 1995), as well as relationship dissatisfaction (Parker et al., 2010). Against this background, the goal of our study is to investigate the role of jealousy as a mediating mechanism in the relationships between partner's smartphone use and corresponding relational outcomes.

The remainder of the paper is organized as follows. In the following section we summarize related work, and derive hypotheses that link *partner's smartphone use* with the feelings of *jealousy* and *relational cohesion* – a critical marker of relational health reflecting "*the degree of togetherness and emotional bonding*" between relational partners (Choi, 2012, p. 92). In the next step, we present results of our qualitative and quantitative studies, based on the responses of "generation Y" smartphone users (aged 26-40). Our qualitative findings suggest that the loss of attention is a key emotional consequence of partner phubbing, providing evidence for the salience of the smartphone-induced jealousy (Bauminger, 2010; Tov-Ruach, 1980). These findings justify further testing of our theoretical model. Implications of our findings for IS research and practitioners are discussed in the concluding section. Our focus on "generation Y" demographic segment has several reasons: First, this age cohort is largely composed of heavy smartphone users, who are most likely to use a wide range of the smartphone's functions (Zickuhr, 2011; Anderson, 2015) and thus might be particularly likely to engage in phubbing. Second, users in the age of 26-40 are more likely to seek meaningful romantic relationships, but at the same time encounter numerous hurdles and ambiguities on their way to do so. Examples include loosing social norms with regard to dating, growing narcissism and unwillingness to compromise characteristic for "generation Y" (Hudson, 2015; Reiner, 2014). Finally, brought up in the 80s and 90s with gadgets and social media still non-existent, generation Y matured into the era of pervasive technology use and are the first 'always-connected' generation (Bull, 2010). Hence, these users might hold conflictual attitudes towards pervasive technologies, when compared to generation Z which is growing with technology as a natural part of their lives (Gardasevic, 2015).

2. Theoretical Background

2.1. Understanding the Concept of Jealousy

Protective in nature, jealousy is typically viewed as a blend of negative feelings, including sadness and worry as well as feelings of exclusion and offense (Schmitt et al. 1994). As such, jealousy is often linked to the *loss of exclusive attention*, with a jealous subject fearing to lose his or her position in the relationship (Bauminger, 2010; Tov-Ruach, 1980). This reaction is natural, since social and romantic relationships universally represent a valuable asset, and hence deserve to be protected (Baumeister and Leary, 1995). While multiple theories have tried to address the antecedents and consequences of jealousy, *the dual factor conceptualization* of jealousy has gained particular importance (Hansen, 1991). According to this approach, emergence and strength of the feelings of jealousy are the product of two contributing factors. On the one hand, a jealous subject should perceive the "*partner's involvement with* *an activity and/or another person as contrary to the definition of relationship*"; on the other hand, the relationship itself should be perceived as valuable (Hansen, 1991, p. 214). While commonly discussed in the context of romantic triads (DeSteno et al., 2006, p. 627), jealousy experience is, hence, not solely limited to them. Instead, activities that are subjectively perceived as threatening, e.g. partner spending too much time at work or with friends, may also antagonize the subject, causing jealous feelings to arise.

Extending this approach, Hansen (1991) additionally introduced the concept of "boundary ambiguity", previously advanced by Boss (1987). Focusing on interactions within families, Boss et al. (1990, p. 5) define boundary ambiguity as "*the family not knowing who is in and who is out of the system*". In other words, "*the family may perceive a physically absent member as psychologically present or may perceive a physically present member as psychologically absent*". Especially the latter form may have a high potential to induce jealousy, as a subject might feel threatened by the psychological absence of the partner – a situation that may run contrary to his or her definition of the relationship. For example, immersion into one's smartphone may result in a boundary ambiguity, with the subject perceiving the other partner as psychologically absent, even though physically present. Facing such painful situation, the subject may try to adopt certain coping strategies. For example, one may try to achieve the psychological presence of the partner, which can be achieved by taking the attempts to change partner's behavior. On the other hand, a strategy aimed to achieve the physical absence of the partner is also possible, with the subject resorting to withdrawal, avoidance or separation (Hansen, 1991). All in all, jealousy is frequently associated with deteriorations in the relationship health (Andersen et al., 1995; Guerrero and Eloy, 1992), as well as an array of other detrimental outcomes oriented towards the self (e.g. reduced self-esteem (Bringle, 1981; Buunk, 1997)), or the target (e.g. violence (Chiffriller and Hennessy, 2007)).

2.2. Understanding the Role of Phubbing in the Relational Context

Past research has shown that all types of interpersonal relationships may be vulnerable to the interference of technology, which can take the form of "*interruptions in face-to-face conversations to the feelings of intrusion an individual experiences*" (McDaniel and Coyne, 2016, p. 85). Owned by 3.4 billion users around the globe (Ericsson Mobility Report, 2015), smartphones may represent the technological phenomenon with the distinct potential to intervene with interpersonal relationships (Billieux, 2012). So far, past research has delivered ambiguous results on the role of smartphones and phubbing in the interpersonal domain. On the one hand, smartphones can be used as a way to connect with others, creating favourable feelings of social connectedness (Chen and Katz, 2009; Devitt and Roker, 2009; Padilla-Walker et al., 2012). For example, serving as a platform for frequent social interaction and exchange of emotional support, smartphones have been shown to promote deeper intimacy between family members (Campbell and Ling, 2009). Furthermore, studies report positive influence of smartphones on the quality of professional communication in healthcare (Wu et al., 2011; Whitlow et al., 2014; Wickersham et al., 2015), on socialization of people with disabilities (O'Neill, 2015) and children suffering from autism (De Leo and Leroy, 2008).

On the other hand, intense engagement with a smartphone inhibits users from fully taking part in their present social surroundings, which may trigger "boundary ambiguity" on the part of others (Hansen, 1991). Indeed, a research report revealed that twenty percent of respondents reported that they could not even remember the phone ever being in a different room than they were (Groarke, 2014). As such, this present absence can be a reason for conflicts in social relationships (Tertadian, 2012; Bernroider et al., 2014), since interpersonal communication is inevitably neglected (Karadag et al., 2015). Furthermore, phubbing has been shown to undermine relational closeness (Przybylski and Weinstein, 2013), since accompanying face-to-face communication is of lower quality and less empathetic (Misra et al., 2014).

In this way smartphones can be seen as a medium that disconnects conversational partners since one might feel left out as the other person is intensively absorbed with his or her smartphone. While any distraction during the time people spend together may provoke negative feelings, past research evidences that not all interrupters are equal, pointing out the stronger feelings of jealousy towards a social object in contrast to an inanimate object like a book (Hart et al., 2004). Perceiving computers to be "fundamentally social" (Nass et al., 2015, p. 72), users develop a strong emotional attachment towards mobile phones and are experiencing "intimacy with their electronic devices" (McDaniel and Coyne, 2016, p. 87 after Turner and Turner, 2013; Vincent et al., 2005; Wehmeyer, 2007). Thus, we believe smartphones are perceived as heavy intruders in communication, leaving the phubbed party feeling not only deprioritized, but also jealous because of the device's extended functionality with social interaction activities as particularly threatening ones. While this undesirable dynamics has been observed across a variety of social contexts, including parental (Radesky et al., 2014), work (Roberts, 2015) and educational (Ling, 2000; Campbell, 2005) settings, recent reports have sent alarming signals regarding the influence of smartphone use on romantic relationships. Often contrasted with friendships, a clear distinction of romantic relationships includes physical attraction, sexuality and a deliberate commitment to long-term, exclusive relationships (Hatfield and Rapson, 1987; Sternberg 1987; Connolly et al., 1999). Specifically, partner phubbing has been linked to lower relationship satisfaction (McDaniel and Coyne, 2016), increased conflict between romantic partners (Coyne et al., 2011; Roberts and David, 2016), and lower well-being (McDaniel and Coyne, 2016; Roberts and David, 2016). Especially partners strongly attached to their significant other are prone to experience conflictual emotions when it comes to the smartphone addiction of the latter (Roberts and David, 2016).

While this dynamics may have far-reaching detrimental implications in the long-run, the mechanisms behind the negative association between partner phubbing and markers of relationship health (e.g. relational cohesion, relationship satisfaction, level of conflict) are still unclear. Considering that partner phubbing inevitably leads to the loss of exclusive attention towards the other party – the core component of the *jealousy* experience (Lazarus, 1991; Tov-Ruach, 1980) - it might be that it is not partner phubbing per se that leads towards relationship dissatisfaction, but rather these are the feelings of *jealousy* this behaviour is triggering that are responsible for this unwanted outcome.

Indeed, while the relationship between partner phubbing and feelings of jealousy has not been explored so far, studies from other related contexts offer solid support for the salience of the jealousy experience in the context of Social Media use (Muscanell et al., 2013; Fox et al., 2014; Tokunaga, 2011; Phillips, 2009) – the focal activity of smartphone users (Smith, 2015; Perez, 2015). For example, the time a partner spends on Facebook has been linked to the heighted feeling of jealousy (Muise et al., 2009). Furthermore, experience of jealousy has been associated with such (somewhat unethical) behaviours, as partner's surveillance (Tokunaga, 2011; Phillips, 2009). Building on these insights, a theoretical model that focuses on the role of jealousy experience as a mechanism in the link between partner's smartphone use and relationship cohesion is developed in the following section.

3. Towards a Theoretical Model

3.1. The Role of Partner Phubbing in Evoking Jealousy

While little scientific evidence is available, initial findings from market research hint at the increasingly important role of smartphones in eliciting jealousy among romantic partners (Waterloo, 2013; E.On Energie Deutschland, 2013). Especially "Generation Y" users may be vulnerable to this threat, since they exhibit high levels of addiction with regard to their smartphone use. For example, such users are likely to exhibit elevated anxiety levels if unable to regularly check their smartphones, reporting to feel "as if a part of them is missing" (Cisco, 2014). Considering their multi-faceted applicability, smartphones may tap into a number of components inherent in the emotional experience of jealousy. First, busy with his or her smartphone, a partner may be unfocused and less responsive with regard to the other party. Experienced in a recurrent pattern, this situation is likely to translate into the perception of "attention loss", which represents one of the core components of jealousy experience (Lazarus, 1991; Ben-Ze'ev, 2010). Moreover, the smartphone can be perceived as a threat to one's exclusive position in the partner's life, which also reflects an important element of the jealousy experience (e.g. Lazarus, 1991; Ben-Ze'ev, 2010; Hart, 2010; Parker et al. 2010; Tov-Ruach, 1980). Additionally, since smartphone use is increasingly associated with the usage of social networking sites, like Facebook, or location-based dating apps (Smith, 2015; Perez, 2015), a partner might fear competition from other parties. Indeed, male users of Facebook – one of the most popular utilities on smartphones (Smith, 2015) – have reported dating as an important reason to join and continue using this site (Bonds-Raacke and Raacke, 2010; Thelwall, 2008). Furthermore, a recent study has shown that smartphones are affecting the dating culture, with 44% of men and 37% of women in the study sample claiming that smartphones make it easier "*to flirt and get to know someone*" (Amplitude Research, 2013). This is in line with the most recent research evidence that suggests that the smartphone-addiction of one's partner can affect interpersonal trust in a negative way and may cause people to put their partner's faithfulness into question (McCormack, 2015) – a common consequence of jealousy (Bevan and Samter, 2004). Taken together, we hypothesize that:

Hypothesis 1 (H1): *The intensity of partner's smartphone use is positively associated with the feelings of jealousy experienced by the other party.*

3.2. The Moderating Role of Personal Smartphone Use

While hypothesis 1 suggests an association between the intensity of partner's smartphone use and the feelings of jealousy, we argue that the strength of this relationship might be moderated by the intensity of the smartphone usage of the significant other. Indeed, the study of Roberts and David (2016) has shown that users who are strongly attached towards their partner are more likely to experience conflict as a result of partner phubbing. Similar outcomes have been observed for the jealousy-induced surveillance behavior, with strongly attached users being more likely to engage in this activity (Fox and Warber, 2014). Moreover, users who themselves use the internet as a leisure time activity appear to be more accepting towards their partner' involvement with phubbing (Klein, 2014). Evidently, partner phubbing is experienced differently when the significant other engages in this activity as well, leading him or her to be more likely to find justification and reasons for this activity. Taken together we argue that:

Hypothesis 1a (H1a): *The relationship between the intensity of partner's smartphone use and feelings of jealousy is moderated by the intensity of the smartphone use by the other party.*

3.3. The Role of Jealousy in Relational Cohesion

Serving to protect romantic bonds (Newberry, 2010), jealousy can in some cases promote more satisfying relationships (Guerrero et al., 1995). Nonetheless, jealousy is often seen as a cause of major relational problems, contributing to aggression and conflict between partners (Guerrero et al., 1995). Indeed, involving a blend of negative emotions, such as anger, sadness, fear and feelings of being hurt and excluded (e.g. Draghi-Lorenz, 2010; Legerstee et al., 2010; Schmitt et al., 1994), jealousy is "a major contributor to relationship dissatisfaction" (Parker et al., 2010, p. 526; Andersen et al., 1995; Bringle et al., 1979) and is predominantly expressed in a negative way. Among others, jealousy can lead to active distancing from the partner (i.e. pulling away from him or her); may involve the jealous subject suffering in silence or displaying such unfavorable emotions as frustration, sadness or anger towards the partner (Bevan and Samter, 2004). Further, giving another the 'silent treatment', sulking, inducing the feelings of guilt (Parker et al., 2010), and being passive aggressive (Adams, 2012) have been identified as common consequences of jealousy experience. Clearly, these expressions threaten to undermine relationship satisfaction, including its related components such as relational cohesion (Spanier, 1976). Indeed, "*broadly defined as the degree of togetherness and emotional bonding*" that relational partners have towards each other (Choi, 2012, p. 92), cohesion is likely to be undermined by the experience of jealousy, as it causes partners to avoid and, consequently, spend less time with each other, thereby interfering with their ability and desire to find time for common activities and conversations (Spanier, 1976). Taken together, we argue that:

Hypotheses 2 (H2): *Feelings of jealousy are negatively associated with perceptions of relational cohesion.*

3.4. The Role of Jealousy as a Mediator

So far, several studies have linked smartphone use with conflict (Tertadian, 2012; McDaniel and Coyne, 2014) and relationship dissatisfaction (McDaniel and Coyne, 2016; Roberts and David, 2016) in romantic relationships. Moreover, additional evidence suggests that the mere presence of a mobile phone can decrease closeness as well as the quality of conversation and connection in dyadic relationships (e.g. Przybylski and Weinstein, 2012). While these findings draw a daunting picture of the future of romance in a smartphone-enabled society at large, little is known about the mechanisms behind these outcomes. Tapping into this critical research question, the study of Klein (2014) illustrates that a high percentage of smartphone-users assume that the usage of one's smartphone in the presence of the other may decrease attention towards that person. Since loss of attention and feelings of exclusivity are at the core of jealousy experience (e.g. Lazarus, 1991; Ben-Ze'ev, 2010; Hart, 2010; Parker et al., 2010; Tov-Ruach, 1980), and jealousy itself is associated with an array of negative relational outcomes, it can be assumed that this is not the usage of the smartphone per se that causes the undesirable outcomes typically attributed to partner phubbing, but these are the feelings of jealousy this usage is evoking, which are responsible for such unwanted relational consequences, as diminishing cohesion between romantic partners. Hence, we hypothesize that:

Hypothesis 3 (H3): *Feelings of jealousy mediate the relationship between the intensity of partner's smartphone use and perceptions of relational cohesion.*

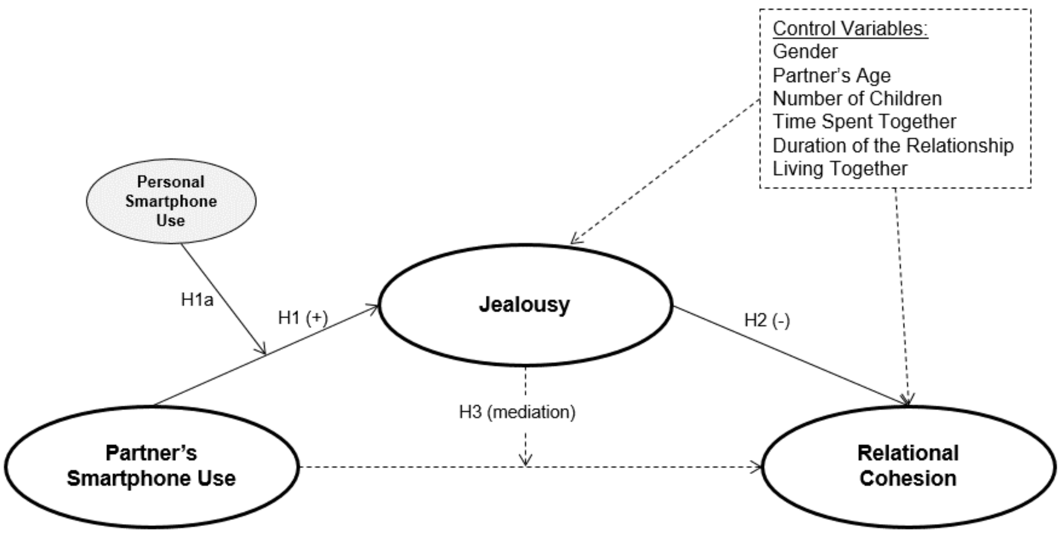

Figure 1. Research model.

Figure 1 summarizes relationships advanced above in a theoretical model. In addition to focal variables, the model includes control variables that have been shown to influence focal constructs in the past research. Specifically, participant gender, partner's age, number of children, time respondent spends with a partner, duration of a relationship, and a living arrangement were included as controls.

4. Methodology and Results

4.1. Survey Design and Flow

To test the advanced hypotheses, a study involving questions for qualitative (here referred to as Study 1) and quantitative (here referred to as Study 2) analysis was conducted. While qualitative questions were included to establish the salience of jealousy feelings in response to partner phubbing (Study 1), scale-based questions posed in Study 2 aimed to explore the relationships proposed in our theoretical model (see Figure 1). Importantly, both studies were presented to the respondents in one online survey. To reduce cognitive overload, questions relating to Study 1 and Study 2 were psychologically separated using a cover story (see Ayyagari et al., 2011).

4.2. Sampling

Respondents were invited to participate in the survey using the mailing list of a large German university and by posting in Facebook groups in the fall of 2015. 40 Amazon.de gift cards (5 Euro value each) were raffled as an incentive to take part in the study. In total, 1475 people completed the survey (completion rate 64.9%). To ensure relevance, observations were cleaned according to the following criteria (resulting in n=1267): 1) a respondent owns a smartphone; 2) a respondent is involved in a romantic relationship; 3) respondent's partner owns a smartphone. Next, 212 observations with a session duration of less than 5 minutes were excluded (mean processing time of the survey comprised 16 minutes and 34 seconds). Finally, considering our focus on the "generation Y", only heterosexual respondents at the age of 26-40 were considered, resulting in a final dataset of 286 observations.

With 64.0%, female respondents are somewhat overrepresented in our sample (male: 36.0%). An overwhelming majority of respondents (79.7%) belongs to the 26-30 age cohort, nearly 17.5% are 31-35 years old and 2.8% of respondents are at the age of 36-40. 76.2% of respondent's partners also belong to generation Y and are 26-40 years old, 18.9% of partners are slightly younger and are 21-25 years old. Approximately 64.7% of respondents have completed their higher education (36.4% have Bachelor and 28.3% have Master Degree). 77.3% of the sample has a student status, 11.9% are employed full-time and 17.8% work part-time. Half of the couples (50.3%) have a common home and 13.6% live "partly" together. Only one respondent claims to have no children, 84.6% of respondents have a child, 7.7% have two children and the rest 7.4% have families with 3 or more children

4.3. Results - Study 1: Exploring Emotions and Reactions Triggered by Partner Phubbing

Considering the lack of studies directly addressing the concept of jealousy in the context of smartphone use, the goal of qualitative questions captured in Study 1 was to explore the salience of the jealousy experience as a reaction to partner phubbing. To achieve this goal, respondents were first asked: "*Think of the last time your partner was using his/her smartphone for too long in your presence. In which situation did it happen?*" Specifying the particular situation (i.e. "the last time") was purposed to decrease the cognitive load and make it easier for a respondent to recall the circumstances and the feelings at that very moment. Assuming that users may experience cases of excessive smartphone use by a partner regularly, this technique allows to reduce the question-answering process by helping the respondent to focus on a particular situation with the highest recall. About one-third of respondents (33.6%) claimed that the incident happened when spending time together at home, 19.6% recalled their partner overusing the smartphone in bed before going to sleep. Further, partner phubbing is noticeable when a couple is having a meal together at home (10.8%), when being on the way in a public transport or in a car (9.8%), and when going out (4.5%). Other occasions were less prominent, with respondents recalling watching TV (2.1%), taking a walk (2.4%), or shopping (0.7%). 22 respondents (8.4%) claimed that their partner has never used the smartphone for too long.

Next, respondents were asked to describe their emotions in this particular situation: "*How have you felt in this regard? Why?*" In total, 252 open answers were provided (34 missing values, correspondingly) and were used for qualitative analyses. Since research does not provide a universal and systematic scheme for coding emotions, inductive theory-driven content analysis was performed by screening the first 100 responses (Russel and Barret, 1999). When sorting, the schematic map of core affect offered by Russel and Barret (1999) was considered since it describes emotions in terms of two consciously accessible elemental processes. The first one - pleasure-displeasure dimension - subjectively summarizes how well a person is doing. The second - activation-deactivation dimension - is related to the level of mobilization or energy. Different possible combinations of two dimensions form a comprehensive set that encompasses all major prototypical emotions (Russel and Barrett, 1999). As a result, the following mutually exclusive seven categories have been identified: 1) *perceived loss of attention*; 2) *anger*; 3) *sadness/suffering*; 4) *boredom*; 5) *neutral/indifferent*; 6) *positive/happiness*; and 7) *other*. In the map of Russel and Barret, *positive/happiness* category would be described by pleasant/active core effect; *anger* as unpleasant/active core effect; *perceived loss of attention*, *sadness/suffering* and *boredom* fall into unpleasant/deactivation quadrant; and *neutral/indifferent* would be placed into the pleasant/deactivation quadrant. Following derived classification scheme (Table 1), 252 responses were coded by two coders independently (coding more than one emotion per response was possible), with Inter-Coder Reliability measured by Krippendorff's Alpha reaching 0.914, which satisfies the threshold of 0.8 (Landis and Koch, 1977). The final decision was taken by consensus. Table 1 presents the summary of the results for the overall sample; and female / male subsamples with a corresponding Wilcoxon rank-sum test used to check for gender-related differences.

Our results suggest that 38.1% of respondents have *neutral* feelings or are *indifferent;* while 4.4% of respondents associate partner phubbing with *positive* emotions. Nonetheless, for the majority of the sample (62.3%) excessive smartphone engagement of a partner was associated with negative jealousy-related feelings. Specifically, 28.6% of the respondents in the overall sample were disturbed by the *loss of partner's attention* – a key element of the jealousy experience (Lazarus, 1991; Ben-Ze'ev, 2010), reporting feeling neglected, unnoticed, less important, turned off, lonely, uninteresting, or isolated, just to name a few. 19.4% felt *angry*, irritated, annoyed, or disturbed amongst other things; and 11.1% of respondents reported feeling *sad* as a result of such behaviour. While only 2 respondents directly described their experience as that of jealousy, the set of negative emotional outcomes provide solid evidence for the salience of jealousy as an emotional reaction to partner phubbing. Indeed, past research has established that anger and sadness are inherent in the experience of jealousy (Bers and Rodin, 1984; Clanton and Smith, 1977); with other authors focusing on the loss of exclusive attention as a key component of jealous feelings (Lazarus, 1991; Ben-Ze'ev, 2010).

Emotion	Key subcategories from open coding	Share of respondents			Wilcoxon test (p>|z|)
		Overall (n=252)	Male (n=90)	Female (n=162)	
Perceived loss of attention	Feeling neglected, unnoticed, less important, turned off, lonely, uninteresting, isolated, rejected, unnecessary, jealous, unconsidered, excluded, dismissed.	28.6%	30.0%	27.8%	0.52
Anger	Feeling irritated, annoyed, disturbed, angry, nervous, under pressure, indignant, displeased, resentful, aggravated.	19.4%	14.4%	22.2%	0.20
Sadness / Suffering	Feeling unhappy, uncomfortable, stupid, unsatisfied, offended, unsure, insecure, worried, bad, not nice, hurt, disrespected, insulted.	11.1%	8.9%	12.3%	0.49
Boredom	Feeling bored.	3.2%	4.4%	2.5%	0.34
Neutral/ indifferent	Feeling ok, no problem, neutral, normal, understanding, indifferent, no matter, unchanged, undisturbed, unaffected, not caring, nothing specific, neither positive nor negative.	38.1%	33.3%	40.7%	0.42
Positive	Feeling good, cool, laugh, super, perfect, glad.	4.4%	7.8%	2.5%	0.04
Other	Feeling curious, tired.	4.8%	6.7%	3.7%	0.24

Table 1. Emotions following partner phubbing.

In the next step, to enhance understanding of the footprint excessive smartphone use leaves on romantic relationships, a follow-up question was posed aiming to elicit coping strategies that are adopted in response to partner phubbing: "*What was your reaction in this situation?" [referring to the situation when the smartphone was overused the last time]*. Supported by the theoretical framework by Hansen (1991), the coding scheme was developed on the basis of Rusbult et al.'s (1986) classification that distinguishes between four types of response to dissatisfaction: exit, voice, loyalty, and neglect (EVLN), and can be described by two primary dimensions: active versus passive, and constructive versus deconstructive. Similar to the previous coding procedure, the first 100 responses were initially screened. For the purpose of precision it was decided to distinguish between the following categories: 1) *voice/intervention*; 2) *voice/curiosity*; 3) *exit/mirror*; 4) *exit/other*; 5) *loyalty*; 6) *feeling negative*; 7) *no reaction*; and 8) *other*. *Voice* measures include expressions of dissatisfaction, with an accompanying attempt to change the situation. Specifically, the category *voice/intervention* subsumes requests to stop using the smartphone; while the category *voice/curiosity* involves such reactions as showing active interest in what is going on in the gadget, e.g. by asking what exactly the partner is doing, who is writing, or looking directly at the partner's smartphone screen. *Exit* strategy implies the dissatisfied person ending the interaction, quitting the partner, or choosing another occupation. We distinguish between the case when a person mirrors the activity of the partner and turns to his or her own smartphone (*exit/mirror*); and when a person pursues another activity beyond the smartphone (*exit/other*). The *loyalty* strategy implies tolerance towards the behaviour of the partner, with a respondent playing a role of passive observer, who does not have an intention to interrupt partner's activity on the smartphone. The category *negative/hurt* summarizes answers that imply some degree of resentment, feelings of being hurt, or annoyance as a result of partner's smartphone overuse. A separate group was created for responses stating *no reaction* at all. In total, 247 answers were coded (39 missing values) from 90 men and 157 female users by two independent coders (coding more than one reaction per response was possible). Resulting Inter-Coder Reliability measured by Krippendorff's Alpha reached 0.727, suggesting an acceptable level of agreement between the coders. The final decision of the code assignment was taken by consensus.

Behavioral strategy	Key subcategories from open coding	Share of respondents			Wilcoxon test (p>|z|)
		Overall (n=248)	Male (n=90)	Female (n=157)	
Voice/ intervention	Active intervention with, or prevention of the smartphone use; making requests to take the smartphone away / stop using it.	27.1%	23.3%	29.3%	0.311
Voice/ curiosity	Expression of clear curiosity; suspicion about the use of the smartphone; looking at the smartphone screen of the partner.	7.3%	5.6%	8.3%	0.429
Exit/ mirror	Reproducing the partner's behaviour, i.e. involvement with one's own smartphone.	6.9%	10.0%	5.1%	0.144
Exit/ other	Choosing another occupation beyond the smartphone.	13.0%	12.2%	13.4%	0.795
Loyalty	Showing patience towards the use of the smartphone by a partner; waiting, understanding, tolerance.	22.3%	28.9%	18.5%	0.059
Feeling negative/ hurt	Feeling offended, insulted; experiencing resentment, annoyance, anger with the situation / partner.	7.3%	7.8%	7.0%	0.823
No reaction	No specific behavioural response	22.3%	20.0%	23.6%	0.518
Other	E.g. not interpretable responses	1.2%	1.1%	1.3%	0.911

Table 2. Reactions following partner phubbing.

We observe that actively intervening with the usage of the smartphone by a partner is the most popular strategy, exercised by 27.1% of the respondents in the overall sample (*voice/intervention*). Next in importance are such strategies as *loyalty* (22.3%) and expressing *no reaction* (22.3%). Interestingly, 13% of the respondents admitted to start doing other things in this situation (*exit/other*), which typically includes watching TV, going to sleep, doing household duties, or reading. At the same time, 6.9% of the respondents copied the smartphone immersion of a partner (*exit/mirror*), suggesting that smartphone use by romantic partners might be contagious and also follow the "tit-for-tat" pattern. Interestingly, such strategy is used by men twice as often as by women, even though this difference is not statistically significant (p-value>0.05, according to Wilcoxon rank-sum (Mann-Whitney) test). *Curiosity was voiced* actively by 7.3% of the respondents who tried to find out what activity their partner was engaged in, who his or her conversational partner was, and what issue it was about. 7.3% of the respondents reported feeling "*negative/hurt*" without implying an active interruption of the partner. All in all, we observe that smartphone overuse provided a rich basis for conflictual situations, with a large share of respondents trying to interfere with this usage or resenting it. As such, the strategies users adopted are typical for the jealousy experience, as described in the past research (Hansen, 1991).

Providing evidence for the prevalence of jealousy as an emotional response to partner phubbing, as well as its conflict-producing nature, qualitative insights obtained in Study 1 provide a solid basis for further quantitative investigation of the role of jealousy in the relationship between partner's use of a smartphone and relational cohesion of partners as a couple (see Figure 1).

4.4. Results - Study 2: Understanding the Role of Jealousy

4.4.1. Survey Design

While we relied on pre-tested measures, where possible, some scales had to be developed new or slightly modified to fit the context of our study. Operationalization of *relational cohesion* was based on a dyadic adjustment scale proposed by Spanier (1976) including the following items: 1) you can calmly discuss something interesting; 2) you laugh together; 3) you exchange thoughts openly with each other; 4) you practice different activities together 5) you find time for each other 6) you are happy in your relationship (1=never; 5=always). To capture *jealousy*, the scale of Schmitt et al. (1994) was adopted, that reflected jealousy as a mix of five emotions: sadness, worry and anger as well as feelings of being excluded and offended. Specifically, respondents were asked to specify "to what extent do you have the following feelings when your partner actively uses the smartphone for too long in your presence?" with items including: 1) it makes me sad; 2) it worries me; 3) I feel excluded; 4) it annoys me; 5) it offends me (1=strongly disagree; 7=strongly agree | "not applicable"). As such, this methodology corresponds to conceptualization of jealousy as a blend of different emotions (Lazarus, 1977; Hansen, 1991). The measure of *partner's smartphone use* was adopted from the cell phone addiction scale of Roberts et al. (2014, p. 256) and included the following items: 1) my partner looks agitated when the smartphone is not in sight; 2) my partner looks nervous when the smartphone battery is almost depleted; 3) my partner spends more and more time on the smartphone; 4) my partner spends more time on the smartphone as he/she should 5) the smartphone is an important part in the life of my partner (1=strongly disagree; 7=strongly agree). Across constructs, the sequence of statements was randomized for each participant. Initially developed in English, the scales were then carefully translated into German. All constructs were measured as reflective. A net sample of 286 observations was included into our analysis (for demographics see section 4.2).

4.4.2. Control Variables

To correctly test the hypothesized relations, several control variables were included into the model. First, considering that emotions are subjective experiences (Barrett, 2006) and the assessment of partner's smartphone usage may depend on one's own behaviour (H1a), *personal smartphone use* was measured by asking "How often do you turn to your smartphone on average per day?" on an 8-point scale: 1= less often than 2 times a day; 8=every 5 minutes (my smartphone is always in my hand). Further, to account for possible bias inherent in a different nature of romantic relationships, we controlled for the *time spent together: "*How much time do you and your partner spend together? (1=practically no time; 6=very much time); whether the couple *lives together* (1=no; 2=partly; 3=yes), *duration of the relationship* (1=less than a year; 6=more than 5 years) and *the number of children* (1=no; 5=more than three). Finally, *respondent's gender* (1=female; 2=male) was included to account for possible differences in gender perceptions; and *partner's age* was controlled for since the latter may be responsible for the so-called "generation gap" - differences of attitudes potentially leading to misunderstanding between people from different age cohorts (VanSlyke, 2003).

4.4.3. Evaluation of the Research Model

Our study is the first to test the relationship between partner phubbing, feelings of jealousy and relational cohesion, which makes our research exploratory in nature. Hence, the partial least squares (PLS) approach was chosen as a method of statistical analysis (Fornell and Bookstein, 1982). Moreover, non-normality of our data and a limited sample size strengthen the case for a variance-based type of evaluation. Hence, SmartPLS 3.0 software was used (Ringle et al., 2015). Evaluation of our research model was done in two steps; the estimation of the Measurement Model (MM) was followed by the assessment of the Structural Model (SM). The MM was evaluated by verifying the criteria for Convergent Validity (CV) and Discriminant Validity (DV). To ensure CV, parameters for Indicator Reliability (IR), Composite Reliability (CR) and Average Variance Extracted (AVE) were assessed. For IR, constructs should explain at least 50 % of the variance of their respective indicators. Items with factor loadings below 0.4 should be removed from the model (Homburg and Giering, 1996). The overwhelming majority of items in all models satisfied the former strict criteria, with most item loadings exceeding the level of 0.7 (Hulland, 1999). Only 4 items measuring *partner's smartphone use* and *relational cohesion* had item loadings closely approximating the required threshold (0.692; 0.685 | 0.691; 0.699). Taken together, IR was assured. Further, CR values for all constructs were higher than the required level of 0.7 (Hulland, 1999), as shown in Table 3. The AVE values for all measured constructs by far surpassed the threshold level of 0.5 (Fornell and Larcker, 1981). Finally, Cronbach's Alpha (CA), a measure of Internal Consistency of construct scales, was higher than the required threshold of 0.7 for all constructs (Nunnally, 1978). Taken together, CV can be assumed. Next, DV was assessed by ensuring that the square root of AVE for each construct was higher than the correlation between this construct and any other construct in the model (Hulland, 1999). This requirement was fulfilled for all constructs in our model. Taken together, our MM is well-specified.

Construct	AVE	CR	CA
Partner's Smartphone Use	0.617	0.889	0.848
Jealousy	0.750	0.937	0.916
Relational Cohesion	0.555	0.882	0.840
Partner's Smartphone Use * Personal Smartphone Use	0.617	0.889	0.871

Table 3. Quality criteria of the latent constructs.

Next, the Structural Model (SM) was assessed as summarized in Figure 2. Significance of path coefficients was determined via a bootstrapping procedure. We find that, *partner's smartphone use* is positively associated with the degree of *jealousy* experienced by the other party (the respondent) *(H1 supported)*. Moreover, the strength of this link is moderated by the *personal smartphone use* of the respondent, with low usage intensity of the respondent associated with heightened *jealousy* perceptions in response to partner's use (*H1a supported*). Furthermore, *jealousy* exerts a significant negative impact on respondent's perceptions of relational cohesion *(H2 supported)*. Among the *six control variables* we tested, only *gender* was associated with the perceptions *of jealousy*, with female users being more jealous in response to partner phubbing than male users.

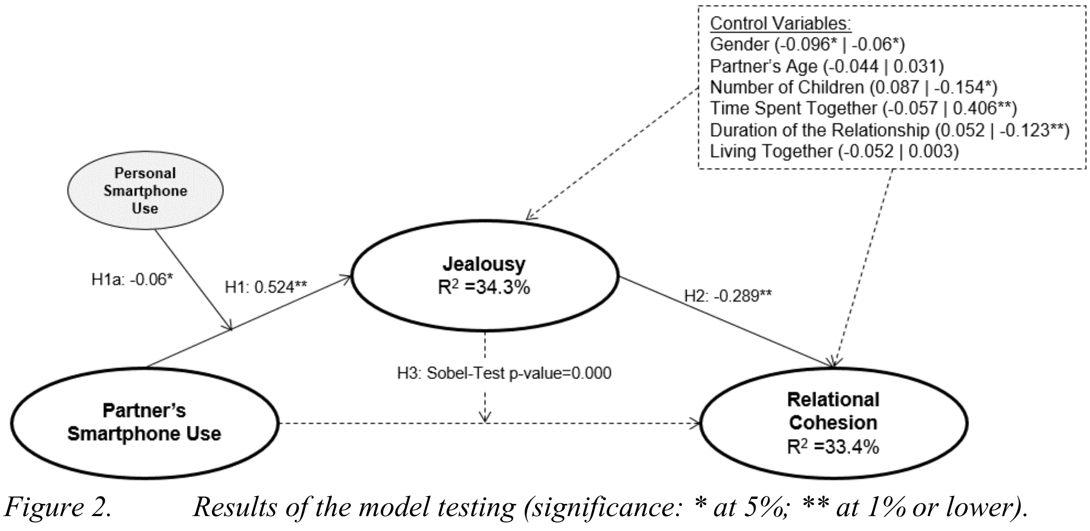

*Figure 2. Results of the model testing (significance: * at 5%; ** at 1% or lower).*

In terms of explanatory power, *jealousy* and six control variables together explain 33.4% of variance in the respondent's perceptions of *relational cohesion* – a noteworthy outcome, considering that a multitude of other factors can strongly influence this construct as well. Overall, this magnitude of explanatory power suggests that smartphone-induced *jealousy* significantly contributes to the relational health of "generation Y" users. For *jealousy,* R^2 has reached 34.3%. Finally, we hypothesized that *jealousy acts as a mediator* between the intensity of *partner's smartphone use* and *relational cohesion*.

To test for this effect, the direct impact of the independent variable – *partner's smartphone use* – on *relational cohesion* was tested first, following (Baron and Keny, 1986). This link was significant and negative *(b= -0.221**)*. However, once the *jealousy* construct was added to the model, the previously significant direct link between *partner's smartphone use* and *relational cohesion* disappeared *(b= -0.071; n.s.)* Furthermore, the *Sobel Test* statistic, typically used to test for mediation, was also significant (*p=0.000*) (Preacher and Leonardelli, 2010-2015). Taken together, we conclude that jealousy *fully mediates* the relationship between *partner's smartphone use* and *relational cohesion (H3 supported)*.

5. Discussion and Managerial Implications

Being an integral part of everyday life for many users, smartphones have the potential to permeate all types of interpersonal settings, including romantic relationships. So far, past research has primarily reported unfavourable consequences of phubbing in the romantic context, establishing smartphones as the cause of conflict (e.g., Roberts and David, 2016), lower relationship satisfaction and reduced well-being (e.g. McDaniel and Coyne, 2016). Contributing to this stream of research, the primary goal of this study was to uncover the mechanism behind this detrimental dynamics. We advance existing theories by proposing and validating a new set of dependences that offer a novel perspective on the undesirable impact of partner phubbing on romantic relationships. We find that observing a partner's smartphone activity may create "boundary ambiguity" (Boss, 1987), leading to heightened feelings of jealousy, which, in turn, may reduce couple's relational cohesion. Moreover, jealousy plays a mediating role in the relationship between partner's smartphone use and relational cohesion, acting as a mechanism behind this undesirable link. Our qualitative results also emphasize the presence and salience of jealousy feelings as a response to partner phubbing. Specifically, "generation Y" respondents report a plethora of negative jealousy-related emotions as a result of their partner's latest phubbing episode (Schmitt, 1994; Tov-Ruach, 1980; Lazarus, 1991), including perceived loss of attention, anger and sadness. As such, our findings challenge a frequently promoted positive view of smartphones as a medium for around-the-clock "connectedness" (Levitas, 2013). In fact, our study draws attention to the often overlooked negative developments, with smartphones impeding emotional bonding and disconnecting partners.

Our findings have implications for IS practitioners including smartphone producers, mobile app providers and other affiliated stakeholders. Indeed, the problem of excessive and, as confirmed by our study, detrimental smartphone use challenges app developers with a need for new innovative solutions. Possible remedies may take the form of an application or special settings, monitoring and managing phubbing activities (Hill, 2015). Moreover, with over 85% of "generation Y" users owning a smartphone (Nielsen, 2014), their impact on users' romantic relationships has meaningful social implications. Since users might be unaware about the ruining impact of phubbing on their romantic relationships, campaigns raising public awareness on this issue might be advisable.

The current study has several limitations. Since most respondents came from Germany, our results are especially valid for countries with a high level of smartphone adoption. Moreover, since partner's smartphone use was measured as a subjective perception of a respondent, future research may apply a more objective assessment of this construct. Further, extending the sample with a broader range of age cohorts may open the opportunity for between-generation comparisons, helping to disentangle psychological mechanisms behind phubbing on a larger scale. Finally, future studies might consider including a social desirability scale to control for the honesty of the responses provided by participants.

References

Adams, S. (2012). “Jealousy in romantic relationships, self-esteem and ego defenses.” Doctoral thesis in clinical psychology, Victoria University. URL: http://vuir.vu.edu.au/21445/1/Sabrina_Adams.pdf (visited on 11/25/2015).

Amplitude Research (2013). *Mobile’s Impact on Dating & Relationships*. URL: http://www.stateofdatingreport.com/docs/FINAL_Mobiles_Impact_on_Dating_and_Relationships_Study.pdf (visited on 11/27/2015).

Anderson, M. (2015). *6 Facts About Americans and Their Smartphones*. *Pew Research Center, April*, *1*. URL: http://www.pewresearch.org/fact-tank/2015/04/01/6-facts-about-americans-and-their-smartphones/ (visited on 04/05/2016).

Anderson, P. A., Eloy, S. V., Guerrero, L. K. and B. H. Spitzberg (1995). Romantic jealousy and relational satisfaction: A look at the impact of jealousy experience and expression. *Communication Reports*, 8, 77-85.

Audience Dialogue (2011). *Know your audience: Chapter 16 – Coding for content analysis*. URL: http://www.audiencedialogue.net/kya16b.html (visited on 11/27/2015).

Ayyagari, R., Grover, V. and R. Purvis (2011). “Technostress: technological antecedents and implications.” *MIS quarterly* 35 (4), 831-858.

Baron, R. M. and D. A. Kenny (1986). “The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations.” *Journal of Personality and Social Psychology* 51, 1173-1182.

Barrett, L. F. (2006). “Are emotions natural kinds?” *Perspectives on psychological science* 1 (1), 28-58.

Baumeister, R. F., and M. R. Leary (1995). The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychological bulletin*, 117(3), 497.

Bauminger, N. (2010). "Jealousy in autism spectrum disorders (ASD)." *Handbook of jealousy: Theory, research, and multidisciplinary approaches*, 267-292.

Beale, R. (2005). “Supporting social interaction with smart phones.” *Pervasive Computing, IEEE* 4 (2), 35-41.

Ben-Ze’ev, A. (2010). Jealousy and romantic love. In: *Handbook of jealousy: Theory, research, and multidisciplinary approaches*. Ed. by S. L. Hart and M. Legerstee. Chichester, West Sussex: Wiley-Blackwell, 40-54.

Bernroider, E. W. N., Krumay, B. and S. Margiol (2014). “Not Without My Smartphone! Impacts of Smartphone Addiction on Smartphone Usage.” In: *Proceedings of the 25th Australasian Conference on Information Systems (ACIS)*, Auckland: New Zealand, 1-20.

Bers, S. A. and J. Rodin (1984). “Social-comparison jealousy: A developmental and motivational study.” *Journal of Personality and Social Psychology* 47 (4), 780-792.

Bevan, J. L. and W. Samter (2004). „Toward a broader conceptualization of jealousy in close relationships: Two exploratory studies.” *Communication Studies* 55 (1), 14-28.

Billieux, J. (2012). “Problematic use of the mobile phone: A literature review and a pathways model.” *Current Psychiatry Reviews* 8 (4), 1-9.

Bonds-Raacke, J. and J. Raacke (2010). “MySpace and Facebook: Identifying dimensions of uses and gratifications for friend networking sites.” *Individual Differences Research* 8 (1), 27-33.

Boss, P (1987). Family stress. *Handbook of marriage and the family*. Springer, US, 695-723.

Boss, P. G., Greenberg, J. R. and D. Pearce-McCall (1990). *Measurement of boundary ambiguity in families*. Minnesota Agricultural Experiment Station, University of Minnesota.

Bringle, R. G and S. Evenbeck (1979). The study of jealousy as a dispositional characteristic. In: *Love and attraction*. Ed. by M. Cook and G. Wilson. New York, NY: Pergamon.

Bringle, R. G. (1981). “Conceptualizing jealousy as a disposition.” *Alternative Lifestyles* 4, 274-290.

Bringle, R., Roach, S., Andier, C. and S. Evenbeck (1979). Measuring the intensity of jealous reactions. *Catalog of Selected Documents in Psychology* 9, 23–24.

Bull, G. (2010). “The always-connected generation.” *Learning and Leading with Technology*, November 2010, 28-29.

Buunk, B. P. (1997). “Personality, birth order and attachment styles as related to various types of jealousy.” *Personality and Individual Differences* 23, 997-1006. doi:10.1016/S0191-8869(97)00136-0

Campbell, M. A. (2005). “The impact of the mobile phone on young people's social life.” In: *Social Change in the 21st Century Conference*. QUT Carseldine: Brisbane.

Campbell, S. W. and R. Ling (2009). Effects of mobile communication. In: *Media effects: Advances in theory and research.* 3rd Edition. Ed. By J. Bryant and M. B. Oliver. New York, NY: Taylor and Francis, 592-606.

Chen, Y. F. and J. E. Katz (2009). „Extending family to school life: College students’ use of the mobile phone.” *International Journal of Human-Computer Studies* 67 (2), 179-191.

Chiffriller, S. H. and J. J. Hennessy (2007). “Male batterer profiles: Support for an empirically generated typology.” *Journal of Offender Rehabilitation* 44, 117-131. doi:10.1300/J076v44n02_05

Choi, A. W. K. (2012). “The relationship between family cohesion and intimacy in dating relationship: A study based on attachment and exchange theory.” *Discovery – SS Student E-Journal* 1, 91-109.
Cisco (2014). *Cisco Connected World Technology Report*. URL: http://www.cisco.com/c/en/us/solutions/enterprise/connected-world-technology-report/index.html#~2012 (visited 11/27/2015).
Clanton, G. and L. G. Smith (1977). Jealousy. Englewood Cliffs. NJ: Prentice-Hall.
Connolly, J., Craig, W., Goldberg, A., and D. Pepler (1999). Conceptions of cross-sex friendships and romantic relationships in early adolescence. *Journal of Youth and Adolescence*, *28*(4), 481-494.
Coyne, S., Stockdale, L., Busby, D., Iverson, B. and D. Grant (2011). “ 'I luv u :)!': A descriptive study of the media use of individuals in romantic relationships.” *Family Relations* 60 (2), 150-162.
Crystal Jiang, L. and J. T. Hancock (2013). “Absence makes the communication grow fonder: Geographic separation, interpersonal media, and intimacy in dating relationships.” *Journal of Communication* 63(3), 556-577.
De Leo, G. and G. Leroy (2008). “Smartphones to facilitate communication and improve social skills of children with severe autism spectrum disorder: special education teachers as proxies.” In: *Proceedings of the 7th international conference on Interaction design and children*, 45-48).
DeSteno, D., Valdesolo, P. and M. Y. Bartlett (2006). "Jealousy and the threatened self: get-ting to the heart of the green-eyed monster." *Journal of personality and social psychology* 91 (4), 626-641.
Devitt, K. and D. Roker (2009). “The role of mobile phones in family communication.” *Children & Society* 23 (3), 189-202.
Dey, A. K., Wac, K., Ferreira, D., Tassini, K., Hong, J. H. and J. Ramos (2011). Getting closer: an empirical investigation of the proximity of user to their smart phones. In: *Proceedings of the 13th international conference on Ubiquitous computing*. ACM. Bejing: China, 163-172.
Draghi-Lorenz, R. (2010). Parental reports of jealousy in early infancy: Growing tension between evidence and theory. In: *Handbook of jealousy: Theory, research, and multidisciplinary approaches*. Ed. by S. L. Hart and M. Legerstee. Chichester, West Sussex: Wiley-Blackwell, 235-266.
Duran, R. L., Kelly, L. and T. Rotaru (2011). “Mobile phones in romantic relationships and the dialectic of autonomy versus connection.” *Communication Quarterly*, 59 (1), 19-36.
E.On Energie Deutschland (2013). *Sind Handys Beziehungskiller? Studie belegt: Deutsche sind eifersüchtig aufs Smartphone*. URL: http://www.presseportal.de/pm/109984/2592438 (visited on 11/05/2015).
Ericsson Mobility Report (2015). *On the Pulse of the Networked Society*. URL: http://www.ericsson.com/res/docs/2015/mobility-report/ericsson-mobility-report-nov-2015.pdf (visited on 11/27/2015).
Fornell, C. and D. F. Larcker (1981). “Structural Equation Models with Unobservable Variables and Measurement Errors.” *Journal of Marketing Research* 18**,** 39-50.
Fornell, C. and F. L. Bookstein (1982). “Two structural equation models: LISREL and PLS applied to consumer exit-voice theory.” *Journal of Marketing Research* 19**,** 440-452.
Fox, J. and K. M. Warber (2014). “Social networking sites in romantic relationships: Attachment, uncertainty, and partner surveillance on Facebook.” *Cyberpsychology, Behavior, and Social Networking* 17 (1), 3-7.
Fox, J., Osborn, J. L. and K. M. Warber (2014). “Relational dialectics and social networking sites: The role of Facebook in romantic relationship escalation, maintenance, conflict, and dissolution.” *Computers in Human Behavior* 35, 527-534.
Gallup (2015). *Most U.S. Smartphone Owners Check Phone at Least Hourly*. URL: http://www.gallup.com/poll/184046/smartphone-owners-check-phone-least-hourly.aspx?utm_source=Economy&utm_medium=newsfeed&utm_campaign=tiles (visited on 27/11/2015).
Gardasevic, S. (2015). Can’t Imagine Your Day Without Your Smartphone? Blame Dopamine! URL: http://domain.me/cant-imagine-day-without-smartphone-blame-dopamine/ (visited on 11/27/2015).
Generation Chart (undated). List of Generation Chart URL: http://www.esds1.pt/site/images/stories/isacosta/secondary_pages/10%C2%BA_block1/Generations%20Chart.pdf (visited on 11/27/2015)
Groarke, H (2014). “The impact of smartphones on social behaviour and relationships.” Department of psychology, Dublin Business School. URL: http://esource.dbs.ie/bitstream/handle/10788/2096/hdip_groake_h_2014.pdf?sequence=1 (visited on 11/24/2015).
Guerrero, L. K. and S. V. Eloy (1992). “Relational satisfaction and jealousy across marital types.” *Communication Reports* 5, 23-31.
Guerrero, L. K., Andersen, P. A., Jorgensen, P. F., Spitzberg, B. H. and S. V. Eloy (1995). „Coping with the green-eyed monster: Conceptualizing and measuring communicative responses to romantic jealousy.” *Western Journal of Communication* 59, 270-304.

Hansen, G. L. (1991). "Jealousy: Its conceptualization, measurement, and integration with family stress theory." The Psychology of Jealousy and Envy. Ed. by P. Salovey. 211-230.
Hart, S. L. (2010). A theoretical model of the development of jealousy: Insight through inquiry into jealousy protest. In: *Handbook of jealousy: Theory, research, and multidisciplinary approaches*. Ed. by S. L. Hart and M. Legerstee. Chichester, West Sussex: Wiley-Blackwell, 331-361.
Hart, S. L. Carrington, H. A., Tronick, E. Z. and S. R. Carrol (2004). "When infants lose exclusive maternal attention: Is it jealousy?." *Infancy* 6 (1), 57-78.
Hatfield, J. and S. Murphy (2007). "The effects of mobile phone use on pedestrian crossing behaviour at signalised and unsignalised intersections." *Accident Analysis & Prevention* 39 (1), 197-205.
Hatfield, E. and R. L. Rapson (1987). Passionate love: New directions in research. *Advances in personal relationships*, 1, 109-139.
Hill, S. (2015). *Is smartphone addiction real? We ask the experts*. URL: http://www.digitaltrends.com/mobile/is-smartphone-addiction-real/ (visited on 11/26/2015).
Homburg, C. and A. Giering (1996). „Konzeptualisierung und Operationalisierung komplexer Konstrukte." *Marketing-Zeitschrift für Forschung und Praxis* 18**,** 5-24.
Hudson, P. (2015). *10 Reasons Why This Generation Is Losing The Ability To Be In Love*. URL: http://elitedaily.com/dating/generation-giving-up-love/989885/ (visited 11/27/2015).
Hulland, J. (1999). "Use of partial least squares (PLS) in strategic management research: A review of four recent studies." *Strategic Management Journal* 20**,** 195-204.
IDC (2013). *How smartphones and social keep us engaged – An IDC research report*. URL: http://www.nu.nl/files/IDC-Facebook%20Always%20Connected%20(1).pdf (visited on 11/26/2015).
Jung, Y. (2013). "What a smartphone is to me: understanding user values in using smartphones." *Information Systems Journal* 24 (4), 299-321.
Karadag, E., Tosuntas, S. B., Erzen, E., Duru, P., Bostan, N., Mizrak, S., Culha, I. and B. Babadag (2015). „Determinants of phubbing, which is the sum of many virtual addictions: A structural equation model." *Journal of Behavioral Addictions*, 4 (2), 60-74.
Klein, V. (2014) Gemeinsam einsam: Phänomen Phubbing. *Information, Wissenschaft & Praxis* 65 (6), 335-340.
Landis J. R. and G. G. Koch (1977). "The measurement of observer agreement for categorical data." *Biometrics* 33 (1), 159-174.
Lazarus, R. S. (1977). "Cognitive and coping processes in emotion." Ed. by A. Monat and R. S. Lazarus. *Stress and coping: an anthology*. Columbia University Press, New York, 145–158.
Lazarus, R. S. (1991). *Emotion and adaption*. New York, NY: Oxford University Press.
Legerstee, M., Ellenbogen, B., Nienhuis, T. and H. Marsh (2010). Social bonds, triadic relationships, and goals: Preconditions for the emergence of human jealousy. In: *Handbook of jealousy: Theory, research, and multidisciplinary approaches*. Ed. by S. L. Hart and M. Legerstee. Chichester, West Sussex: Wiley-Blackwell, 163-191.
Levitas, D. (2013). *Always Connected: How Smartphones And Social Keep Us Engaged*. URL: https://www.academia.edu/7042070/Always_Connected_How_Smartphones_And_Social_Keep_Us_Engaged (visited on 11/23/2015).
Ling, R. (2000). "The impact of the mobile telephone on four established social institutions." In: *ISSEI2000 Conference of the International Society for the Study of European Ideas*. Bergen: Norway, 14-18.
Lyndon A., Bonds-Raacke J. and A. D. Cratty (2011). "College students' Facebook stalking of ex-partners." *Cyberpsychology, Behavior, & Social Networking* 14, 711–716.
McCormack, M. (2015). *The role of smartphones and technology in sexual romantic lives*. Project Report, Durham University.
McDaniel, B. T. and S. M. Coyne (2016). "Technoference": The interference of technology in couple relationships and implications for women's personal and relational well-being. *Psychology of Popular Media Culture*, *5*(1), 85-98.
Misra, S., Cheng, L., Genevie, J. and M. Yuan (2014). "The iphone effect the quality of in-person social interactions in the presence of mobile devices." *Environment and Behavior*.
Mod, G. B. B. (2010). "Reading romance: The impact Facebook rituals can have on a romantic relationship." *Journal of comparative research in anthropology and sociology*, (2), 61-77.
Mohan, P., Padmanabhan, V. N. and R. Ramjee (2008). "Nericell: rich monitoring of road and traffic conditions using mobile smartphones." In: *Proceedings of the 6th ACM conference on Embedded network sensor systems*, 323-336.
Muise, A., Christofides, E. and S. Desmarais (2009). "More information than you ever wanted: Does Facebook bring out the green-eyed monster of jealousy?" *CyberPsychology & Behavior*, 12(4), 441-444.
Muscanell, N. L., Guadagno, R. E., Rice, L. and S. Murphy (2013). "Don't it make my brown eyes green? An analysis of Facebook use and romantic jealousy." *Cyberpsychology, Behavior, and Social Networking* 16 (4), 237-242.

Nass, C., Steuer, J. and Tauber, E. R. (2015). “Computers are social actors.” *Human Factors in Computing Systems*, 72-78.
Newberry, M. A. (2010). “The positive and negative effects of jealousy on relationship quality: A meta-analysis.” UNF Theses and Dissertations, Paper 280, University of North Florida. URL: http://digitalcommons.unf.edu/etd/380/ (visited on 11/25/2015).
Nielsen (2014). *Mobile Millennials: Over 85% of Generation Y Owns Smartphones*. URL: http://www.nielsen.com/us/en/insights/news/2014/mobile-millennials-over-85-percent-of-generation-y-owns-smartphones.html (visited on 11/27/2015).
Nunnally, J. C. (1978). *Psychometric theory*. McGraw-Hill, NY: New York.
O'Neill, M. (2015). Smartphone Apps for People with Disabilities. URL: http://www.makeuseof.com/tag/5-smartphone-apps-people-disabilities/ (visited on 11/20/2015).
Padilla-Walker, L. M., Coyne, S. M. and A. M. Fraser (2012). “Getting a High-Speed Family Connection: Associations Between Family Media Use and Family Connection.” *Family Relations* 61 (3), 426-440.
Parker, J. G., Kruse, S. A. and J. Wargo Aikinis (2010). When friends have other friends: Friendship jealousy in childhood and early adolescence. In: *Handbook of jealousy: Theory, research, and multidisciplinary approaches*. Ed. by S. L. Hart and M. Legerstee. Chichester, West Sussex: Wiley-Blackwell, 516-546.
Perez, S. (2015). *Consumers spend 85% of time on smartphones in apps, but only 5 apps see heavy use*. URL: http://techcrunch.com/2015/06/22/consumers-spend-85-of-time-on-smartphones-in-apps-but-only-5-apps-see-heavy-use/#.hdibm0j:5bJy (visited on 11/27/2015).
PewInternet (2011), Generations and their gadgets, URL: http://www.pewinternet.org/files/old-media/Files/Reports/2011 /PIP_Generations_and_Gadgets.pdf [Created: 2011-02-03; Accessed: 2015-12-15].
Phillips, M. (2009). MySpace or yours?: Social networking sites surveillance in romantic relationships. Paper presented at the annual meeting of the Western States Communication Association, Mesa, AZ.
Phys.org (2015). *US smartphone use surges, at expense of other gadgets (Update)*. URL: http://phys.org/news/2015-10-smartphone-center-tech-universe-survey.html (visited on 27/11/2015).
Preacher, K. J. and G. J. Leonardelli (2010-2015). Calculation for the Sobel test. URL: http://quantpsy.org/sobel/sobel.htm (visited on 11/27/2015).
Przybylski, A., K. and N. Weinstein (2012). „Can you connect with me now? How the presence of mobile communication technology influences face-to-face conversation quality.” *Journal of Social and Personal Relationships* 30 (3), 217-246.
Radesky, J. S., Kistin, C. J., Zuckerman, B., Nitzberg, K., Gross, J., Kaplan-Sanoff, M., Augustyn, M. and M. Silverstein (2014). „Patterns of mobile device use by caregivers and children during meals in fast food restaurants.” *Pediatrics* 133 (4), e843-e849.
Reiner, A. (2014). *Love, Actually: Teaching Generation Y the Basics of a Strong.* URL. Relationshiphttp://www.nytimes.com/2014/02/09/education/edlife/teaching-generation-y-the-basics-of-a-strong-relationship.html (visited on 11/27/2015).
Ringle, C. M., Wende, S. and J.-M. Becker (2015). SmartPLS 3. Bönningstedt: SmartPLS. URL: http://www.smartpls.com (visited on 11/27/2015).
Roberts, J. A. (2015). *Did you get phubbed by your boss?.* URL: http://edition.cnn.com/2015/10/06/opinions/roberts-boss-phubbing/ (visited on 11/27/2015).
Roberts, J. A. and M. E. David (2016). “My life has become a major distraction from my cell phone: Partner phubbing and relationship satisfaction among romantic partners.” *Computers in Human Behavior*, 54, 134-141.
Roberts, J., Yaya, L. and C. Manolis (2014). “The invisible addiction: Cell-phone activities and addiction among male and female college students.” *Journal of Behavioral Addictions 3* (4), 254-265.
Rusbult, C. E., Johnson, D. J. and G. D. Morrow (1986). “Determinants and consequences of exit, voice, loyalty, and neglect: Responses to dissatisfaction in adult romantic in-volvements.” *Human Relations* 39, 45-63.
Russell, J. A. and L. F. Barrett (1999). “Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant.” *Journal of personality and social psychology* 76 (5), 805-819.
Salehan, M. and A. Negahban (2013). “Social networking on smartphones: When mobile phones become addictive.” *Computers in Human Behavior* 29, 2632-2639.
Schmitt, M. J., Montada, L. and K. Falkenau (1994). „*Modellierung der generalisierten und bereichsspezifischen Eifersuchsneigung mittels Strukturgleichungen*.“ Berichte aus der Arbeitsgruppe „Verantworung, Gerechtigkeit, Moral“, Nr. 071. URL: http://www.gerechtigkeitsforschung.de/berichte/beri071.pdf (visited on 11/24/2015).
Smith, A. (2015). *Chapter Three: A „week in the life analysis of smartphone users.* URL: http://www.pewinternet.org/2015/04/01/chapter-three-a-week-in-the-life-analysis-of-smartphone-users/ (visited on 11/27/2015).
Spanier, G. B. (1976) "Measuring dyadic adjustment: New scales for assessing the quality of marriage and similar dyads." *Journal of Marriage and the Family*, 15-28.

Sternberg, R. J. (1987). Liking versus loving: A comparative evaluation of theories. *Psychological Bulletin*, 102(3), 331.
Tertadian, E. (2012). "The Smart Phone as a Conflicting Third Party in Interpersonal Relationships." *Communication Studies Undergraduate Publications: Presentations and Projects,* Paper 14.
Thelwall, M. (2008). "Social networks, gender, and friending: An analysis of mySpace member profiles." *Journal of the American Society for Information Science and Technology* 59 (8), 1321-1330.
Theiss, J. A. and D. H. Solomon (2006). A relational turbulence model of communication about irritations in romantic relationships. *Communication Research*, 33(5), 391-418.
Tokunaga, R. S. (2011). "Social networking site or social surveillance site? Understanding the use of interpersonal electronic surveillance in romantic relationships." *Computers in Human Behavior* 27 (2), 705-713.
Toma, C. L. and M. Choi (2016). Mobile Media Matters: Media Use and Relationship Satisfaction among Geographically Close Dating Couples. URL: https://www.academia.edu/16752532/Mobile_media_matters_Media_use_and_relationship_satisfaction_among_geographically_collocated_dating_couples (visited on 11/26/2015).
Tov-Ruach, L. (1980). When friends have other friends: Friendship jealousy in childhood and early adolescence. In: *Handbook of jealousy: Theory, research, and multidisciplinary approaches*. Ed. by A. Oksenberg Rorty. London: University of California Press, 465-488.
Turner, P. and Turner, S. (2013). Emotional and aesthetic attachment to digital artefacts, in: Cognition, Technology and Work, 15, 403-414.
Utz, S. and C. J. Beukeboom (2011). "The role of social network sites in romantic relationships: Effects on jealousy and relationship happiness." J*ournal of Computer-Mediated Communication* 16 (4), 511-527.
VanSlyke, T. (2003). "Digital natives, digital immigrants: Some thoughts from the generation gap." *The technology source 7* (3).
Vincent, J., Haddon, L. and L. Hamill (2005). The influence of mobile phone users on the design of 3G products and services, in: *Journal of the Communication Network*, 4, 69-73.
Waterloo, L. (2013). *You love your iPhone more than me! Almost half of cheaters admit they strayed because partner paid more attention to their mobile*. URL: http://www.dailymail.co.uk/femail/article-2500833/You-love-iphone-Almost-half-cheaters-admit-strayed-partner-paid-attention-mobile.html (visited on 11/25/2015).
Wehmeyer, K. (2007). Assessing users' attachment to their mobile devices, in: Management of Mobile Businesses, Proceedings of the ICMB, Toronto 2007; New York: IEEE Computer Society.
White, J., Thompson, C., Turner, H., Dougherty, B. and D. C. Schmidt (2011). "Wreckwatch: Automatic traffic accident detection and notification with smartphones." *Mobile Networks and Applications* 16 (3), 285-303.
Whitlow, M. L., Drake, E., Tullmann, D., Hoke, G. and D. Barth (2014). "Bringing technology to the bedside: Using smartphones to improve interprofessional communication." *Computers Informatics Nursing* 32 (7), 305-311.
Wickersham, A., Salehi, H., Pennathur, P., Linson, E., Are, C. and A. Kamath (2015). "Smartphones in the Improvement of Inter-Professional Communication." *Journal of Hospital Medicine* 10.
Woollaston, V. (2013). *How often do you check your phone? The average person does it 110 times a DAY (and up to every 6 seconds in the evening)*. URL: http://www.dailymail.co.uk/sciencetech/article-2449632/How-check-phone-The-average-person-does-110-times-DAY-6-seconds-evening.html (visited on 11/27/2015).
Wu, R., Rossos, P., Quan, S., Reeves, S., Lo, V., Wong, B., Cheung, M. and D. Morra (2011). "An evaluation of the use of smartphones to communicate between clinicians: a mixed-methods study." *Journal of medical Internet research* 13 (3).
Zickuhr, K. (2011). Generations and their gadgets. Pew Internet & American Life Project, 20, URL: http://www.pewinternet.org/2011/02/03/generations-and-their-gadgets/ (visited on 27/11/2015).

Erklärung zum Promotionsvorhaben

Hiermit erkläre ich, dass ich zuvor noch keiner Promotionsprüfung unterzogen wurde sowie ich mich noch um keine Zulassung an der Humboldt-Universität zu Berlin bzw. einer anderen Universität beworben habe. Weiterhin habe ich noch keiner Universität oder ähnlichen Einrichtung eine Dissertation vorgelegt.

Annika Baumann

Eidesstattliche Versicherung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne fremde Hilfe nur unter Verwendung der angeführten Literatur angefertigt habe.

Ich bezeuge durch meine Unterschrift, dass meine Angaben über die bei der Abfassung meiner Dissertation benutzten Hilfsmittel, über die mir zuteil gewordene Hilfe sowie über frühere Begutachtungen meiner Dissertation in jeder Hinsicht der Wahrheit entsprechen.

Annika Baumann